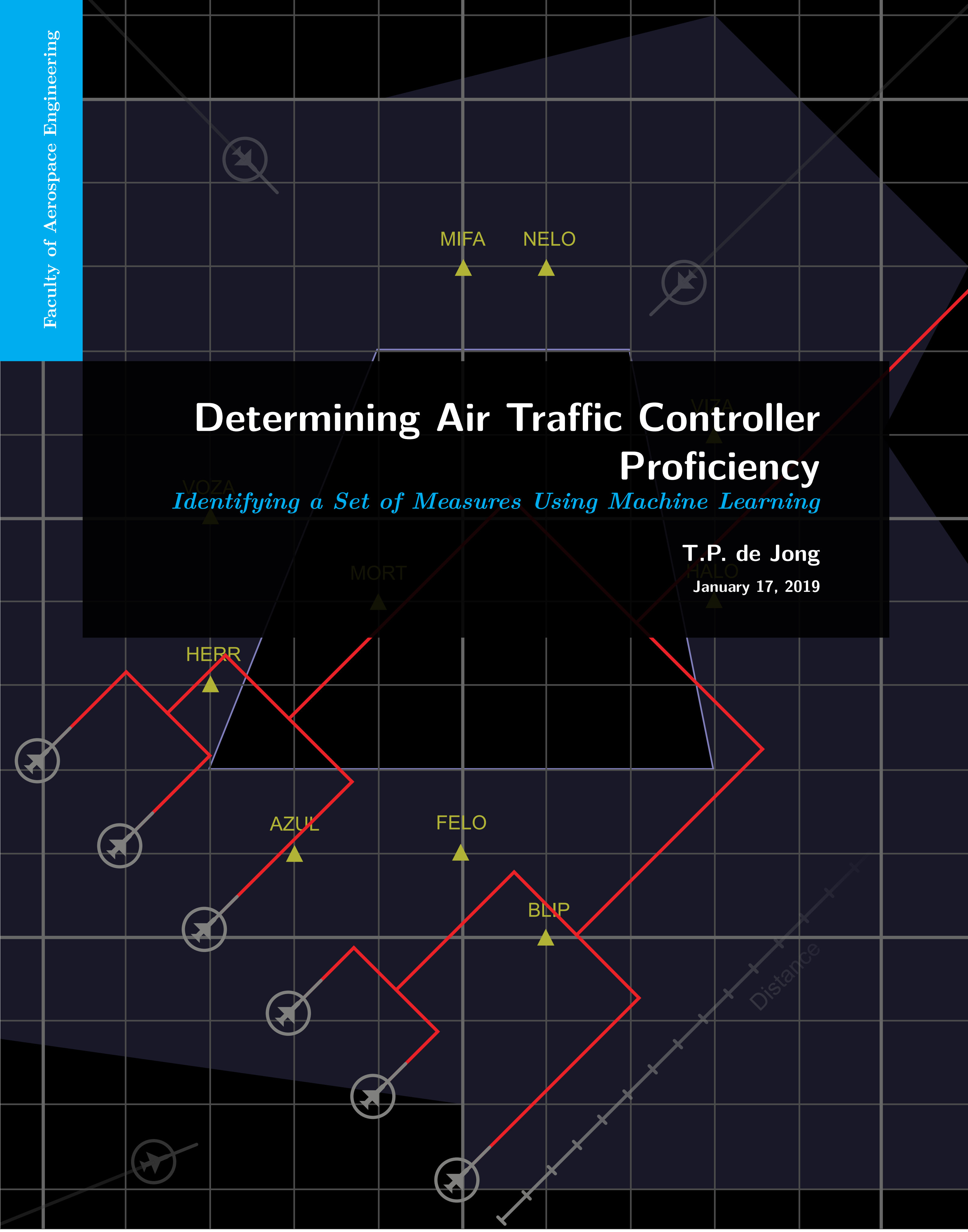
Faculty of Aerospace Engineering



Determining Air Traffic Controller Proficiency

Identifying a Set of Measures Using Machine Learning

T.P. de Jong

January 17, 2019

Delft University of Technology

Challenge the future

Determining Air Traffic Controller Proficiency

Identifying a Set of Measures Using Machine Learning

Master of Science Thesis

For obtaining the degree of Master of Science in Aerospace Engineering
at Delft University of Technology

T.P. de Jong

January 17, 2019

Faculty of Aerospace Engineering · Delft University of Technology

Delft University of Technology

Delft University Of Technology
Department Of
Control and Simulation

The undersigned hereby certify that they have read and recommend to the Faculty of Aerospace Engineering for acceptance a thesis entitled **"Determining Air Traffic Controller Proficiency"** by **T.P. de Jong** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: January 17, 2019

Readers:

Dr.ir. C. Borst

Ir. Y. B. Eisma

Dr.ir. M. M. van Paassen

Prof.dr.ir. M. Mulder

Dr. O. A. Sharpanskykh

Preface

This report is the result of a full year of work towards finishing my MSc. A full year of new experiences and challenges which were sometimes hard, but especially enjoyable to do. In this report, entitled ”Determining Air Traffic Controller Proficiency - Identifying a Set of Measures Using Machine Learning”, I present to you the work I have performed in order to obtain my master’s degree at the department Control & Simulation at the Faculty of Aerospace Engineering of the Delft University of Technology.

I would like to express my graduate to Clark, who has been my daily supervisor during the whole project. The (almost) weekly meetings were very valuable to the progress of the project and without the brainstorms, the feedback and the enthusiasm the performed work did not have the quality as presented in this report. I would also like to thank Yke Bauke who offered great support during the preliminary phase of this project and continuing phase after it. I would like to thank Max and René for the supervision during the meetings we had. Your feedback gave me new helpful insights and suggestions during the process.

Throughout this project I held many interesting conversations about this project with friends and fellow students. This gave me many new interesting ideas and motivation. I want to express my graduate towards my friends and fellow students not only for the support during this project, but also for the memorable time I had in Delft.

Thank you to my girlfriend Ellen, who has been very supportive in my daily life during the entirety of my master. With her love she brightens up my day. Finally, I would like to thank my parents, brothers and sister who have always been very interested and supportive in the decisions I made during the entirety of my bachelor and master.

Tjitte de Jong
January 17, 2019

Acronyms

ACC	Area Control Center
ACoPOS	ATC Cognitive Process & Operational Situation model
AI	Artificial Intelligence
ATC	Air Traffic Control
ATCo	Air Traffic Controller
BSS	Between-cluster Sum of Squares
CPA	Closest Point of Approach
CTA	Control Area
DCPA	Distance at Closest Point of Approach
DCT	Direct
EFL	Executive Flight Level
HDG	Heading
ICAO	International Civil Aviation Organization
LOS	Loss of Separation
ML	Machine Learning
PCA	Principal Component Analysis
PZ	Protected Zone
R/T	Radio Telephony
SPD	Speed
SSD	Solution Space Diagram
TCPA	Time to Closest Point of Approach
TSS	Total Sum of Squares
WSS	Within-cluster Sum of Squares

List of Figures

List of Tables

Contents

Part I

Master of Science Thesis Paper

Determining Air Traffic Controller Proficiency: Identifying a Set of Measures Using Machine Learning

T.P. de Jong

Supervisors: C. Borst, Y.B. Eisma, M.M. van Paassen and M. Mulder
Delft University of Technology, Faculty of Aerospace Engineering, Delft, The Netherlands
t.p.dejong@mail.com, {c.borst, m.m.vanpaassen, m.mulder}@tudelft.nl

***Abstract*—A high drop-out rate is present during current-day air traffic controller (ATCo) training, because the required expertise level is not reached. The determination of the expertise level of ATCo students is currently performed using subjective assessments at a late stage in the training by means of high-fidelity simulator sessions. It is desired to objectively measure expertise earlier and more frequently in training to monitor the progress of the student. However, it is currently unknown which objective measures might describe the expertise level of an ATCo. This paper presents a method that identifies a set of objective measures that can classify an ATCo's expertise level using a genetic algorithm and hierarchical agglomerative clustering. A large set of possible objective measures and a dataset containing data from 10 ATCos (intermediate and pro level) is used. The method found a set of 8 measures that can cluster the 10 ATCo's in the two expertise groups very accurately (97,5% accuracy). The genetic algorithm showed a preference for measures that have a distinction in the results between the expertise groups. However, not all selected measures show a difference between the expertise groups, resulting in signs of overfitting. Furthermore, these measures only provided limited feedback for the individual ATCos. Clustering the results of the 10 ATCo's gave valuable information about the overall expertise level of an ATCo, as compared to the average intermediate- or pro-ATCo.**

***Index Terms*—Air traffic control, proficiency, expertise level, machine learning, genetic algorithm, hierarchical agglomerative clustering**

I. Introduction

The selection of candidates for the air traffic controller (ATCo) training is strict, because of the high demanding nature of the job. Due to the highly complex and dynamic environment, ATCos need to process large amounts of dynamically changing information while maintaining a good balance between safety and efficiency within environmental constraints defined to the procedures in his or her sector [1]. These candidates need to acquire the required competences during the limited training period (usually two or three years). Unfortunately, even after a strict selection, a high drop-out rate is present during the training period, because the required expertise level is not reached or cannot be reached within the training period [2]. This is undesirable, because a large amount of effort has been put into these students.

The determination of the level of expertise is currently done by using subjective assessments. The instructor assesses the level of the students based on his or her own experience. These subjective assessments are performed at a late stage in the training by means of high-fidelity simulator sessions. It is more preferable to objectively measure expertise earlier and more frequently in training, to continuously monitor the progress of the student. However, it is currently unknown which objective measures might describe the expertise level of an ATCo. The determination of this combination of objective measures is researched in this paper.

To find a good combination of objective measures, a genetic algorithm is used with machine learning to determine the performance of this combination. Studies that use machine learning in order to determine experience level have already been performed in other fields, such as speaking proficiency levels [3], billiard players [4] and surgeons [5]. However, no prior research has been conducted using machine learning to find a set of objective measures that could determine the expertise level of ATCos.

The research objective of this thesis is to identify a set of objective measures that can classify an air traffic controller's level of control expertise by using machine learning techniques.

In this research a large set of objective measures is created from prior research. The dataset used to test the set of measures consists of data from four air traffic scenarios solved by ten different ATCos. Four participants were retired ATCos and six participants completed a multiple day extensive ATC-course and/or had worked as a researcher in the ATC field [6]. To search in a large set of measures a genetic algorithm will be used to find the best subset. To determine the performance of the subset of measures the ATCo data are clustered using a hierarchical clustering algorithm. The accuracy and stability of the clusters can be calculated from the results to determine a single performance value. The genetic algorithm uses this performance value to continue its search for the best set of measures.

In the ideal case this best set of measures describes the difference between novice-, intermediate- and pro-ATCos. ATCo students could be objectively assessed for this best set of measures during training. By doing this on a frequent base the progress towards pro-ATCo behavior can be monitored. This could reveal the competences that are still underdeveloped, so that more attention and guidance can be given to these competences during training [7]. Eventually, this could lead to a more effective effort and a lower the amount of drop-outs.

This paper starts with the theoretical motivation (Section II) in which the ATCo competences, ATC structural elements and corresponding metrics are discussed. Section III provides a description of the ATCo data and how this data are obtained. Section IV describes the methodology used to find the set of measures that best describes the different ATCo expertise groups. This section includes the selected measures from the theoretical motivation, the transformation of the dataset and how the best set is obtained from the selected measures using the processed dataset. The results from the genetic algorithm and clustering algorithm are shown in Section V. Section VI describes the method and results of a sensitivity analysis. A sensitivity analysis reveals the robustness of the best set of measures and the clusters, and reveals which measures and ATCos contribute the most to the accuracy of the clusters. A discussion of the results and the sensitivity analysis is discussed in Section VII. Finally, the conclusion of this research is given in Section VIII.

II. Theoretical motivation

The main goal of an ATCo is to ensure a safe, orderly and expeditious flow of air traffic in his or her sector [8]. The expertise level is determined by a set of ATC related competences. High requirements are set for these competences because of the cognitive complexity for the ATCo. This cognitive complexity cannot be seen separately from the operational situation [1]. Therefore, both the ATCo competences and the ATC structural elements in the operational situation are discussed.

A. Competences and Structural Elements

Schuver-van Blanken et al. developed the ATCo Cognitive Process & Operational Situation (ACoPOS) model which provides a competence-based training model with elements of the operational air traffic control (ATC) situation (Figure 1) [1]. This model also shows the relationship between the ATCo competences and ATC elements.

The blocks *situation assessment*, *problem solving & decision making*, and *attention management & workload management* form a representation of the cognitive processes of the ATCos. The competences present in the cognitive processes can only be assessed subjectively. However, the result of the cognitive process is reflected in the performed actions, which can be objectively assessed [8].

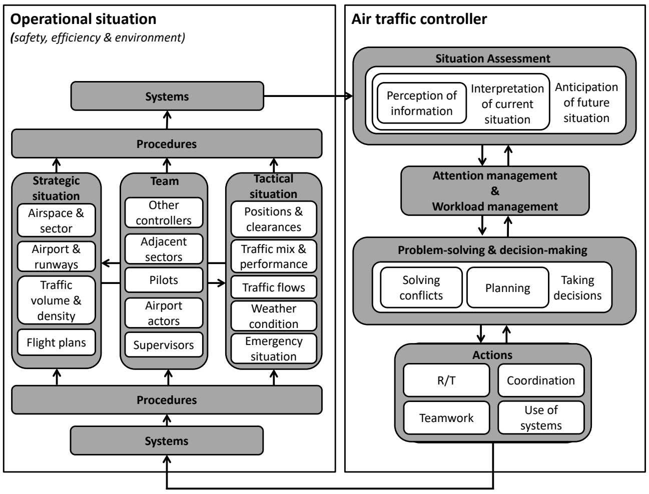


Fig. 1: The ATCo Cognitive Process & Operational Situation model (ACoPOS model) (adapted from Schuver-van Blanken et al. [1]).

In the *Action* block, Radio telephony (R/T) focuses on the interaction between the ATCo and the aircraft, while the other competences focus on interaction between ATCos or interaction between the ATCo and the equipment [7].

Finding a correct balance between safety, efficiency and environmental constraints is the core task of the ATCo [1]. Safety and efficiency are determined by the actions of the ATCo. To what extent safety and efficiency are reached can be determined by the flight movements and R/T recordings. Therefore, safety and efficiency are highly linked to the ATCo competences.

In the ACoPOS model, the *strategic situation* determines the physical boundaries in which an ATCo needs to handle traffic [1]. The *tactical situation* is marked by the dynamic and changing nature of the situation. This aspect has an effect on the cognitive complexity for the ATCo, since the results of the changing situations are often non-preferable or unexpected [1].

The elements present in the block *Team* represent the influence of other controllers, adjacent sectors, pilots, airport sectors or supervisors. Since this research solely focuses on the expertise of a single ATCo and not his or her team-working capabilities, the elements in this block will not be included in this research.

How the traffic needs to be handled formally inside the sector is determined by the procedures [1]. These procedures can limit the ATCo's action degrees of freedom. Therefore, the element *procedures* has an influence on the ATCo.

B. Metrics

In order to determine to what extent the ATCo competences are reflected, metrics are linked to each ATCo competence and ATC structural element. Table I shows selected competences and elements from the ACoPOS model with the linked metrics.

TABLE I: Metrics to assess ATCo competences and ATC structural elements.

	R/T	
1	Consistency in the type of instructions	Kallus et al. [9]
2	The way of R/T use to keep the workload low	Hilburn [10]
	Safety	
3	Use sufficient safety buffers	Schuver-van Blanken and van Merriënboer [11]
	Efficiency	
4	Maximize efficiency	Oprins et al. [8], Kirwan and Flynn [12]
5	Moment of traffic handling	(From hypothesis)
6	Create an expeditious flow of traffic	ICAO [13]
	Traffic volume & density	
7	Availability of solution space	Schuver-van Blanken and Roerdink [2]
	Procedures	
8	Variability in procedures	Schuver-van Blanken et al. [1]

According to Kallus et al., ATCos have an internal *"conflict solution library"* [9, p.46]. The most frequent and commonly used solutions come first in mind. These solutions need certain types of instructions. It is therefore expected that experienced ATCos are more consistent in the use of certain types of instructions. Furthermore, an experienced ATCo must keep his or her workload as low as possible [10]. The way R/T is used has an influence on this workload. ATCos could guide aircraft in such a way that traffic flows require more monitoring, which increases workload. Furthermore, communicating more with the aircraft takes extra time. Therefore, the way R/T is used to keep the workload low can be used as a metric to determine experience.

A way to ensure safety is to be more conservative or cautious, depending on the ATCo's age and fatigue, the experienced workload, or factors like bad weather [14]. Therefore, sufficient safety buffers need to be maintained to cope with uncertainties or to become more cautious [11]. Another way to assess safety is to ensure that separation minimums are maintained [13] and that the amount of errors in the used procedures is minimized [1].

By interviewing ATCos, Kirwan and Flynn found many principles and strategies used by ATCos [12]. One of those principles is to minimize the additional track miles flown. A metric that is related to the minimization of the additional track miles flown is to minimize the delay time of the aircraft [8]. When an aircraft needs to fly additional track miles, it is possible that a delay will occur, unless the ATCo allows the aircraft to fly faster. Both minimizing additional track miles as minimizing delay time are part of the maximization of efficiency.

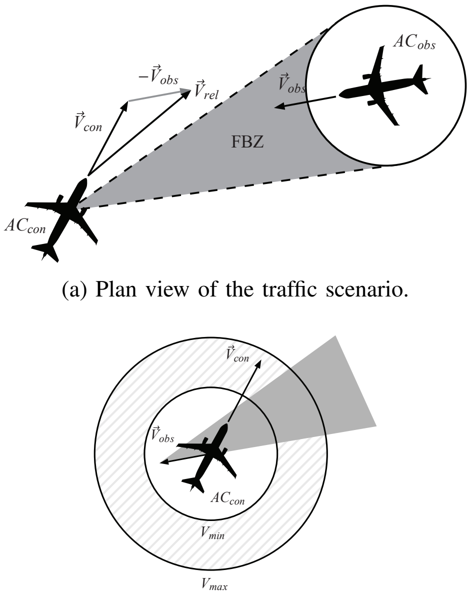
Since an experienced ATCo has more controller experience compared to a novice ATCo, it is reasonable to think that the experienced ATCo has a quicker overview of the situation and handles traffic quicker. Therefore, it is reasonable to think that an experienced ATCo will give all level, heading or speed changes far before the aircraft leaves the sector. The moment of traffic handling can therefore be seen as an efficiency metric. Furthermore, part of the task of an ATCo is to create an expeditious flow of air traffic in his or her sector [13]. A higher outflow of aircraft might indicate a higher efficiency.

To determine the effect of the *traffic volume & density* on the ATCo, the availability of the solution space needs to be determined. According to the findings of Schuver-van Blanken and Roerdink, ATCos create solution space or use the solution space that is already available [2]. A possibility to operationalize solution space is by using a Solution Space Diagram (SSD) which is developed for purposes such as workload determination [15], decision-making support [16] and airspace complexity [17]. The SSD can be graphically represented as shown in Figure 2. This 2D SSD covers all possible heading/velocity combinations in which the aircraft can safely move within the sector and all possible heading/velocity combinations in which the aircraft is on a conflict course with another aircraft [18]. The area of the SSD in which the aircraft is on a conflict course with another aircraft is the occupied SSD area. This area is represented by the dark grey area within the V_{min} and V_{max} bounds in Figure 2. The availability of the solution space has influence on the ability to use certain conflict resolution strategies (like lateral resolutions), the efficiency and the prevention of future problems [2].

Within a sector, a procedure can result in several options for the ATCo. For example, in the AMS ACC South Sector aircraft need to be transferred to Schiphol Approach at an initial approach fix (IAF) between flight level 70 and 100 [19]. This means that there is a variability in this procedure. How this variability is used could express differences between the ATCo expertise groups. Deviating from the procedures does not indicate a lesser expertise level, but could actually indicate a higher expertise level, because this could be performed to resolve certain conflicts or emergency situations. Therefore, it is more interesting to look at the consistency in the variability in the use of procedures.

III. Data description

The dataset used in this research consists of data from four air traffic scenarios solved by ten different ATCos. Four participants were retired ATCos (the professional group)



(a) Plan view of the traffic scenario.

(b) Solution Space Diagram for the controlled aircraft. The gray striped SSD area is the available solution space. The triangular dark gray SSD area bounded by V_{min} and V_{max} is the occupied SSD area.

Fig. 2: Solution Space Diagram area of the controlled aircraft (adapted from Mercado Velasco et al.).

and six participants completed a multiple day extensive ATC-course and/or had worked as a researcher in the ATC field (the ATC course/research group) [6]. The experience of the professional group ranged from 33 to 35 years of experience. Two pro-ATCos were area control center ATCos and two were tower approach control ATCos.

The dataset was obtained in an experiment from Somers [6]. The goal of this experiment was *"to investigate the correlation of the 3D solution space metric with the workload"* [6]. A simplified, medium-fidelity, three-dimensional simulator, based on the Amsterdam Advanced Air Traffic Control (AAA) system used in the Netherlands, was used which showed a sector comparable to the AMS ACC South Sector (Figure 3) [6]. The participant could control the traffic using a separate control window, which could be operated by using a mouse or a touchscreen (Figure 4). The traffic was controlled by clicking on the aircraft and then giving a command using the command window. The aircraft were separated by giving heading, level or speed commands.

A few simplifications were made compared to the actual AAA system, to minimize training effects and to solely test the workload caused by the traffic in the sector [6]. There were three aircraft categories which were shown in the aircraft label: light, medium or heavy. A caution that a loss of separation will occur within 120 seconds was made visible by changing the aircraft color to orange. An actual loss of separation within 60 seconds changed the aircraft color to red. Commands given by the command window were always followed by the aircraft immediately. All aircraft had the same 5 NM protected zone. Furthermore, an option was present to turn the protected zone circle and the speed vector on/off to aid the participants separate the traffic. However, it was not logged in the data whether the protected zone circle and speed vector were on or off. Only the change in on/off was recorded.

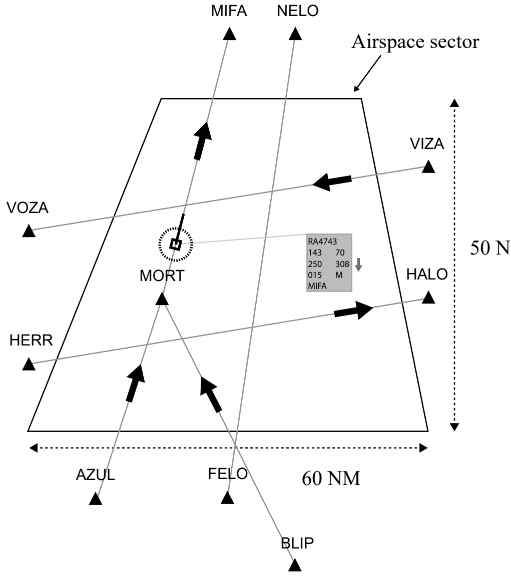


Fig. 3: The three-dimensional simulator screen with the different waypoints, routes and one aircraft visible in the middle (Adapted from [6]).

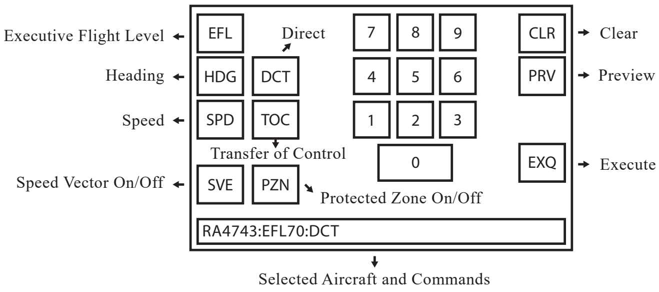


Fig. 4: The command window participants had to use to control the aircraft (Adapted from [6]).

Taking into account the simplifications, this means that many competences and structural elements discussed in Section II do not emerge in the logged data. These include the cognitive process of the ATCo, teamwork, coordination, use of systems, team influences, tactical situation influences, system influences, and a large part of the strategic situation influences. Competences and structural elements that do emerge are safety, efficiency, R/T, procedures and traffic volume & density. Since the use of voice commands was replaced by the command window, R/T was measured by the input in the command window.

The task of the controllers was to separate the traffic and hand them over to the adjacent sectors at predefined flight levels [6]. Before the aircraft left the sector, a transfer of control had to be given. The participant had to follow the following specific instructions [6]:

- Inbound traffic coming from AZUL and BLIP and going to the northern waypoint MIFA, has to be merged and leave the sector between FL 70 - 100.
- Outbound traffic from NELO to FELO has to leave the sector at F200.
- Over flights towards HALO have to be handed over at FL210.
- Over flights towards VOZA leave the sector at the same flight level as they enter (FL140).
- Aircraft have to be given a transfer of control before they leave the sector.
- When aircraft are given a transfer of control they have to be separated (at least 1000 ft vertically and 5NM horizontally) from each other and should not be involved in any conflicts.

The ATCos needed the solve 4 different scenarios. The differences between the scenarios was characterized by the amount of traffic, traffic mix, traffic merges, overtakes, crossings and deviating aircraft [6]. In general, the differences ranged from high/many to low/few. Each scenario had a duration of 20 minutes.

The obtained data consist of two files per ATCo and scenario. One file contains the given commands to the aircraft from the command window, including a timestamp in seconds. The other file contains the data from the simulation window (Figure 3). This file includes, per logpoint, among others, the aircraft position, (commanded) flight level, (commanded) heading and (commanded) speed. A logpoint was recorded every 3 seconds during the experiment. With 10 ATCos, this resulted in 40 radar logs and 40 command logs.

IV. Methodology

This section describes the methodology that is used to find a set of measures that best describes the different ATCo expertise groups of which the results are shown in the next Section (Section V). This measure selection process is shown in Figure 5.

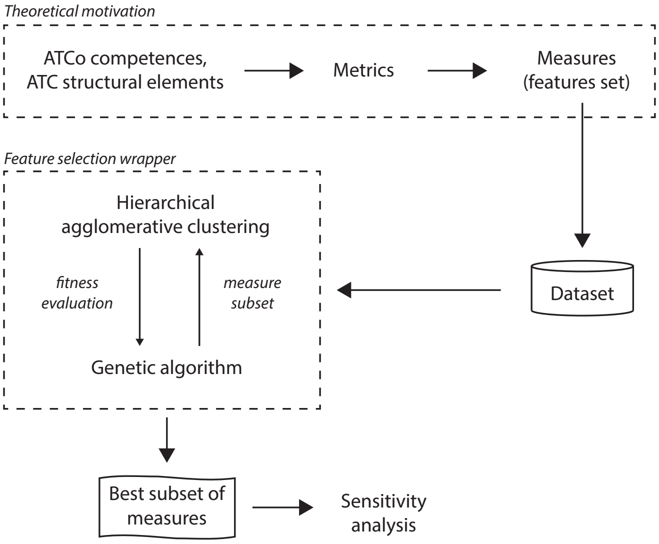


Fig. 5: Flow diagram of the measure selection process.

A. Measures

A total of 59 measures, linked to the metrics in Table I, are used in which a set of measures can be extracted by the genetic algorithm (Table II). These 59 measures are gathered per ATCo and scenario. Table III shows the structure of the processed data. The 40 rows represent the data from each of the 4 scenarios (scenario *a* to *d*) solved by 10 ATCos. C1 to C6 represents the course-group. P1 to P4 represents the pro-group. The 59 columns represent the data of the 59 measures. The majority of these measures are measures of central tendency (like the mean) or measures of variability (like the standard deviation, maximum value and minimum value). Also, ratios, summations and mean squared errors (MSE) are used. This is done to summarize the generated data of each individual ATCo.

TABLE II: Measures corresponding to the metrics of the ATCo competences and ATC structural elements.

	R/T	**Measures**
1	Consistency in the type of instructions	Number of DCT, EFL, HDG and SPD commands
2	The way of R/T use to keep the workload low	Total number of commands; Amount of level changes per aircraft
	Safety	**Measures**
3	Use sufficient safety buffers	Relative distance between aircraft; Average TCPA, DCPA and TLOS
	Efficiency	**Measures**
4	Maximize efficiency	Spent time in sector; Amount of aircraft that reached their waypoint
5	Moment of traffic handling	Trackpenalty when using level, heading or speed commands
6	Create an expeditious flow of traffic	Outflow of traffic in the sector
	Traffic volume & density	**Measures**
7	Availability of solution space	Mean occupied SSD area at each given command; Total occupied SSD area of every aircraft in the sector
	Procedures	**Measures**
8	Variability in procedures	Altitude of aircraft leaving the sector

Looking at the type of instructions given by ATCos it is

TABLE III: The structure of the processed data. The 40 rows represent the data from each of the 4 scenarios (scenario *a* to *d*) solved by 10 ATCos. C1 to C6 represents the course-group. P1 to P4 represents the pro-group. The 59 columns represent the data of the 59 measures.

	Measure 1	Measure 2	···	Measure 59
C1a			···	
C1b			···	
C1c			···	
C1d			···	
C2a			···	
⋮	⋮	⋮	⋮	⋮
C6d			···	
P1a			···	
⋮	⋮	⋮	⋮	⋮
P4d			···	

expected that more experienced ATCos are more consistent in the type of given instructions. To measure this, the number of direct (DCT), executive flight level (EFL), heading (HDG) and speed (SPD) commands are gathered. The ratio between each of these commands and the sum of these commands are used as measures, resulting in 4 different measures.

Since ATCos prefer to use level changes over heading changes when the workload is high [20], and want to keep the workload low by minimizing the number of instructions [10], the amount of level changes per aircraft is stored. From these values the mean, standard deviation and maximum value are used as measures, resulting in 3 measures. Furthermore, the total number of instructions is used as a measure.

To measure the use of sufficient safety buffers the relative distances between the aircraft is stored per logpoint. Only aircraft are considered that do not have reached their destination waypoint and have not been issued with a TOC command, so that the ATCo still has an influence on the aircraft. From these values the mean, standard deviation, minimum value and the maximum value per logpoint is stored. Then again from these new values the mean and standard deviation are calculated, resulting in 8 measures. Furthermore, from the mean values of the relative distances between the aircraft per logpoint the maximum and minimum value are calculated, resulting in 2 measures.

Other values to measure the use of sufficient safety buffers is using the time to closest point of approach (TCPA), distance at closest point of approach (DCPA) and the time to loss of separation (TLOS). Only the positive TCPA, DCPA and TLOS values are considered. The averages from all aircraft of these values are stored per logpoint and again only of all aircraft that the ATCo can control at that logpoint. From these averages the mean, standard deviation, maximum value and minimum value are used as measures, resulting in 12 measures.

To achieve an efficient flow of traffic it is expected that aircraft spend as little time as possible inside the sector. Therefore, the time that an aircraft spends inside the sector is stored. Furthermore, the number of aircraft that reached their destination waypoint is stored per logpoint. From these values the mean and the standard deviation are used as measures, resulting in 4 measures.

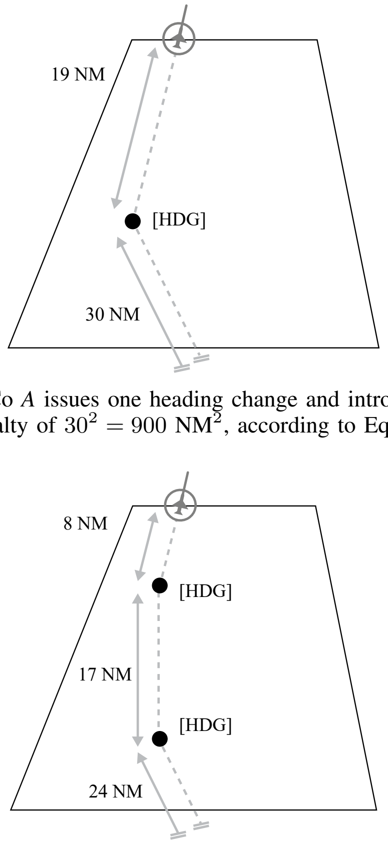
Considering the moment of traffic handling by ATCos it is expected that more experienced ATCos will give all level, heading or speed changes far before the aircraft leaves the sector. To measure to what extent this metric is expressed, for each aircraft the sum of the squared track miles when a level, heading or speed command is given is used (Equation 1). Figure 6 shows an example of this trackpenalty when using heading commands for a single aircraft. For each aircraft and each command type the sum of the squared track miles, when a particular command type is given, is obtained. These sums are taken together to get a single sum of squared track miles for each command type. This results in a trackpenalty when using level, heading or speed commands. The ratio between these trackpenalties and the sum of these trackpenalties are used as measures, resulting in 4 measures.

$$\text{Trackpenalty: } \sum\nolimits_{\substack{\text{squared track miles} \\ \text{at command}}} = a^2 + b^2 + ... \qquad (1)$$

a Track miles of aircraft at first command [NM]
b Track miles of aircraft at second command [NM]

Furthermore, two sets of values of the flow of aircraft flying out of the sector (the outflow) are stored. The first set contains per logpoint the number of finished aircraft divided by the total number of logpoints of the scenario. A finished aircraft is an aircraft that reached its destination waypoint. From this set the mean, standard deviation and maximum value are used as measures, resulting in 3 measures. Furthermore, the maximum value and the MSE of the change in this outflow are used as measures, resulting in 2 measures. The second set contains per logpoint the number of finished aircraft divided by the current logpoint. From this set the mean and standard deviation are used as measures, resulting in 2 measures.

According to the findings of Schuver-van Blanken and Roerdink, ATCos create solution space or use the solution space that is already available [2]. To assess this metric the occupied SSD area is used. This occupied area is represented as a ratio of the total SSD area with a value between 0 and 1. The total SSD area is the total annular area bounded by V_{min} and V_{max} as shown in Figure 2. The mean occupied SSD area of every aircraft in the sector at each given command could be used to see to what extend the solution space changes between the given commands. The change between these mean occupied SSD areas is calculated. From the changes



(a) ATCo *A* issues one heading change and introduces a trackpenalty of $30^2 = 900$ NM^2, according to Equation 1.

(b) ATCo *B* issues two heading changes and introduces a trackpenalty of $24^2 + (24 + 17)^2 = 2257$ NM^2, according to Equation 1.

Fig. 6: ATCo *A* handles the aircraft far before the aircraft leaves the sector compared to ATCo *B*. This is represented by a lower trackpenalty for ATCo *A* compared to ATCo *B*.

the percentage difference, mean, standard deviation and the ratio between an increase or a decrease are used as measures. Furthermore, the mean and the standard deviation of the mean occupied SSD areas are used as measures. This results in 6 measures.

The occupied SSD area of every aircraft in the sector at each logpoint is stored. From these values the mean, standard deviation, maximum value and minimum value are used as measures, resulting in 4 measures.

In the sector the ATCos had the option to let the aircraft leave the sector between flight level 70 and 100 for aircraft flying to waypoint MIFA. Therefore, the flight level of aircraft flying to waypoint MIFA when leaving the sector is stored. From these values the mean, standard deviation, maximum value and minimum value are used as measures, resulting in 4 measures.

B. Input Dataset

Table III shows the structure of the processed data which is used as the input dataset for the genetic algorithm. After the 59 measures are extracted from the data it is important to remove highly correlated measures ($|\rho| > 0.99$). The reason is that the size of the measure set can introduce problems in clustering. These problems include a large computation time, difficulty in interpretation of results and the introduction of the curse of dimensionality [21]. The curse of dimensionality describes that when the dimensionality is high enough, the distance between the nearest points is no different from that of other points [22]. Furthermore, by removing highly correlated measures redundant measures are also removed. Irrelevant measures do not provide any useful information to the clustering method and can even negatively impact the clustering results [23]. However, since the measures are based on theoretical motivation it is not expected that irrelevant measures are present. After removing the highly correlated measures, 55 measures remained. This results in an input dataset with a size of 40x55.

Figure 7 shows the process of splitting the input dataset and standardizing the training set and test set. First, the input dataset is split row-wise into 80% training data (5 course-ATCos and 3 pro-ATCos) and 20% test data (1 course-ATCo and 1 pro-ATCo). This is performed to minimize overfitting: when the model is trained too specialized on a dataset. A smaller training set than 80% results in a set with too few samples to represent the complete ATCo population. The training set consist of the processed data from each scenario of ATCo *C1*, *C3*, *C4*, *C5*, *C6*, *P2*, *P3* and *P4*. The test set consist of the processed data from each scenario of ATCo *C2* and *P1*. Therefore, the training set and the test set have a matrix size of 32x55 and 8x55, respectively.

$$\begin{gathered} x'_{n,m} = \frac{x_{n,m} - \mu_m}{\sigma_m} \\ n \in \mathbb{Z} : n \in [1, 32] \\ m \in \mathbb{Z} : m \in [1, 55] \end{gathered} \tag{2}$$

All values in the training set are standardized according to Equation 2 to get all the measures on the same scale as the training set with zero mean and unit variance. In this equation $x'_{n,m}$ represents the standardized value of ATCo scenario n and measure m, $x_{n,m}$ the original value, μ_m the mean of measure m and σ_m the standard deviation of measure m. In other words: the value in each column in the matrix is standardized using the mean and standard deviation of the corresponding column. The values for μ_m and σ_m of the training set are used to standardize the values of the test set. This is done because the genetic algorithm and clustering algorithm uses the scale of the training set to search the best set of measures.

The 55 measures can be represented as a 55-bit binary array. When a measure is part of the set, it is represented as

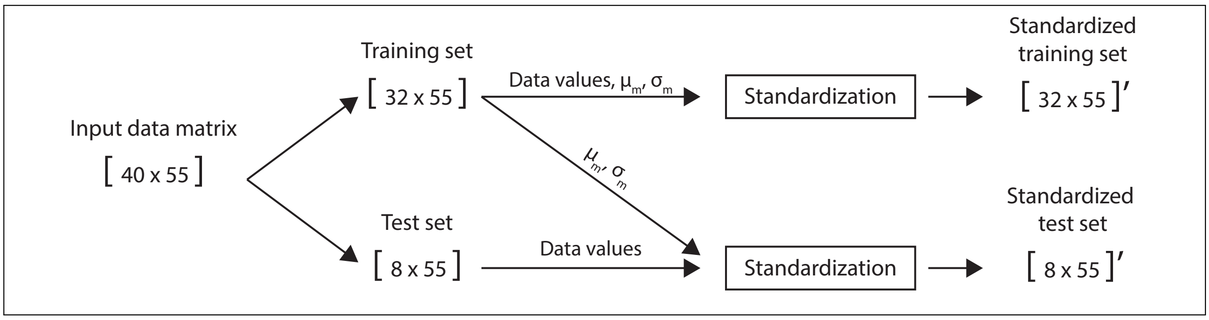


Fig. 7: The process of splitting the input dataset and standardizing the training set and test set.

TABLE IV: The structure of the population of sets of measures. For each set of measures a fitness value can be calculated to determine the performance.

	Measure 1	Measure 2	...	Measure 55	
Set 1	0	0	...	0	→ Fitness value
Set 2	1	0	...	0	→ Fitness value
Set 3	1	1	...	1	→ Fitness value
:	:	:	...	:	
Set 100	1	1	...	0	→ Fitness value

a 1. It is represented as a 0 when a measure is not part of the set. By initializing 100 possible set of measures randomly, a population of sets of measures can be made (Table IV).

C. Feature Selection Wrapper

Since the number of samples (ATCos and scenarios) is small, it is desired to directly get information about the relationship between all the samples. Therefore, hierarchical clustering is used which organizes the data in a hierarchical structure according to the distance matrix [21]. Within hierarchical clustering there are two methods of clustering: agglomerative and divisive. Agglomerative clustering is a "bottom up" method which starts with *N* clusters containing a single data object each [21]. In the process that follows the individual clusters are merged which finally leads to one single cluster. Divisive clustering is a "top down" method that starts as a single cluster containing all the data [21]. In the process that follows the clusters are divided until there are only clusters containing a single data object. Looking at agglomerative clustering, the computational complexity is at least $\mathcal{O}(n^2)$. For divisive clustering the complexity is even worse with a computational complexity of $\mathcal{O}(2^n)$ [21]. Therefore, agglomerative clustering will be used in this research.

The difference between different agglomerative clustering algorithms is determined by the linkage criterion which determines the distance between clusters based on the definition of the distance [21]. Ward's method is a linkage criterion that tries to keep the total within-cluster sum of squares at a minimum value [21]. It is desired to create clusters containing ATCos with similar experience level and that the ATCos within each cluster are close to each other. Therefore, total within-cluster sum of squares should be minimized. Since Ward's method already tries to keep the total within-cluster sum of squares at a minimum value, using this linkage criterion could lead to good clustering results. The Euclidean distance measure is used in general when using Ward's method.

Genetic algorithms are defined as *"a class of stochastic search algorithms based on biological evolution"* [24, p.222]. A genetic algorithm measures the performance of the individual set of measures based on a fitness function to carry out reproduction. When reproduction takes place, crossover and mutation take place. After a number of successive reproductions, the result is that lower performing set of measures will disappear and higher performing set will excel [24]. A genetic algorithm is used to search the best subset of measures from the pre-selected measures relatively quick. A genetic algorithm does not get stuck in a local optimum, because it uses mutation which is equivalent to a random search in the search domain [24].

The feature selection wrapper selects from the pre-selected measures a subset of measures which leads to the most distinct clusters describing the different ATCo expertise groups. First, an initial subset from the pre-selected measures is constructed. The wrapper loop uses hierarchical agglomerative clustering to cluster the subset data and evaluates the performance of the formed clusters. This performance is used as a fitness criterion for the genetic algorithm. Based on the fitness evaluation, the genetic algorithm creates a new subset of measures which is, again, used as an input for the clustering algorithm to determine the performance of this subset. The genetic algorithm creates new subsets based on the current best performing subset and when the new subset is better than the current best performing subset, the new subset becomes the best performing subset. After a stop criterion is reached, the output of the wrapper is the best performing subset of measures.

Figure 8 shows the process of obtaining the best set of measures using a random initial population (Table IV), the

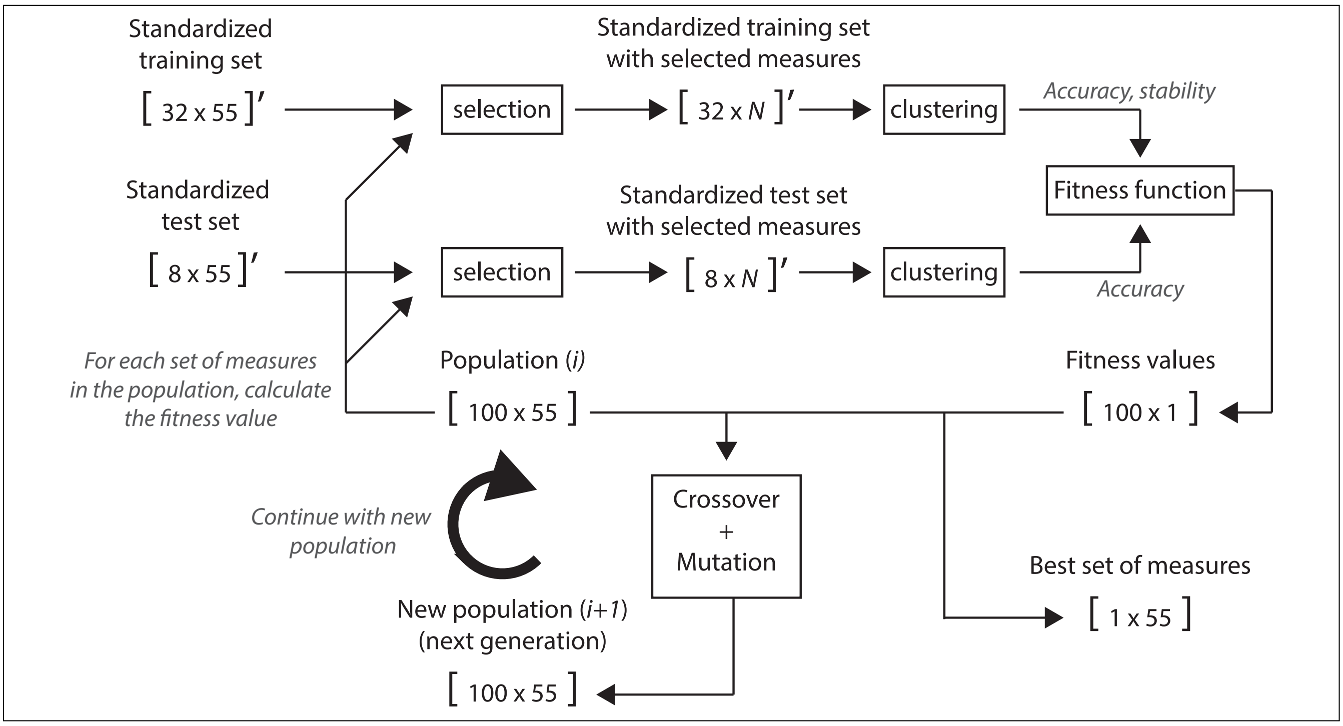


Fig. 8: Iterative process of finding the best set of measures.

standardized training set, and the standardized test set. First, for each set of measures in the population the fitness is calculated. This is done by selection of the measures from the standardized training and test set according to the measures present in the set of measures from the population. The column size *N* depends on the amount of measures in a set of measures. Hierarchical agglomerative clustering is performed on both sets with selected measures.

From the clustering results the clustering accuracy is calculated of both sets using a variant of the confusion matrix. A confusion matrix is a matrix that shows the predicted classification and the actual classification of an object [25]. The accuracy is the proportion of the total number of predictions that were correct [25]. A variant of the confusion matrix is used (Table V), because it is not clear beforehand what the types of the predicted clusters are: a course-cluster or a pro-cluster. This means that two accuracies can be calculated using the two true positives (TP) and two true negatives (TN). The highest accuracy of both accuracies determines the accuracy of the clustering results (Equation 3) and therefore also what the types are of the predicted clusters. An accuracy ranges from not accurate (0) to perfectly accurate (1).

$$Accuracy = max(\frac{TP1+TN1}{TP1+TP2+TN1+TN2}, \frac{TP2+TN2}{TP1+TP2+TN1+TN2}) \tag{3}$$

TABLE V: The variant of the confusion matrix that is used to calculate the maximum accuracy.

		Actual	
		Course	Pro
Predicted	Course or Pro	TP1	TN2
	Pro or Course	TP2	TN1

From the clustering results of the standardized training set with selected measures the stability is calculated. This is only performed for the training set, because the selection of the measures is mainly based on the training set. The stability is calculated using non-parametric bootstrapping [26], because it is not known what the distribution is of the data. A wrongly assumed distribution in parametric bootstrapping can result in wrong stability results. The bootstrapping algorithm uses a similarity measure called the Jaccard coefficient. The Jaccard similarity between two clusters is the ratio between the number of elements in the intersection of the clusters and the elements in the union of the clusters. The Jaccard coefficient ranges from no similarity (0) to perfect similarity (1). Unstable clusters will result in a Jaccard coefficient of 0.6 or less. Lesser stable clusters result in a coefficient between 0.6 and 0.75. Highly stable clusters will result in a coefficient of 0.85 or more [26]. The following steps are used to get the stability of the clusters [26]:

1) Cluster the training set
2) Resample the original training set with the same size of

the original training set. Alter the data in the resampled set by creating duplicates of ATCos in the set and/or removing ATCos.

3) Cluster the new resampled dataset
4) Calculate the similarity between the clusters of the original training set and the clusters of the resampled set. This will result in a Jaccard coefficient for each cluster. Store these coefficients for this run.
5) Repeat from step 2 N times
6) Calculate the average Jaccard coefficient for each cluster over the N runs

The two accuracy values (one of the training set and one of the test set) and two stability values (of both clusters from the training set) are used in the fitness function to determine the performance of the set of measures. Both accuracies are equally important, because it is desired that the correct experience level is assigned to the ATCos in both seen (training) and unseen (test) data. Also, both stabilities are equally important, because both original course- and pro-cluster must still exist when the data are altered.

The accuracies are more important than the stabilities, because it is the goal of this research that the set of measures can accurately determine the experience level of the ATCos. Since the accuracy and stability both range from 0 to 1, a weight of 10 is given to the accuracy to make it more important than the stability and to keep the value of the accuracy most likely on the left side and the stability on the right side of the decimal separator. This results in a single relation between the accuracies and stabilities that produces a single fitness value (Equation 4). This fitness value has a range from 0 to 11.

$$\begin{aligned} fitness = 10 * min(ACC1, ACC2) + \\ min(Stab1, Stab2) \end{aligned} \tag{4}$$

$ACC1$ Accuracy training set [0 $\sim$ 1]
$ACC2$ Accuracy test set [0 $\sim$ 1]
$Stab1$ Stability training set course-cluster [0 $\sim$ 1]
$Stab2$ Stability training set pro-cluster [0 $\sim$ 1]

The fitness value of all 100 sets of measures in the population are calculated creating an array of fitness values with a size of 100x1. With use of the fitness values, the set of measures from the population with the highest fitness is the current best set of measures. The iteration continues by creating a new population from the current population using crossover and mutation. The sets of measures with a high fitness have a high probability to create new offspring for the new population. This new generation is then used in the next iteration as the population.

There is no specific rule which sets the probability when crossover and mutation will occur (p_c and p_m) during the generation of a new population. According to Negnevitsky typical values for p_c and p_m are 0.7 and 0.001, respectively [24]. Since the search domain is large, it is desired to rely on the good *"genes"* of the previous population. Therefore, a value of 0.8 will be used for the crossover probability. When population fitness converges, it is desired to rely more on the mutation of *"genes"* to escape a local optimum. The typical mutation probability is very low for a 55-bit binary array representation when there is a high reliance on bit-flips in the array during convergence. Therefore, a much higher mutation probability of 0.4 is used. This mutation probability is still lower than the crossover probability, because the good *"genes"* of the previous population must not be lost in the next generation.

D. Sensitivity Analysis

The best performing subset is subjected to a sensitivity analysis. A sensitivity analysis does not only reveal the robustness of the subset and the clusters, but can also reveal which measures and ATCos contribute the most to the accuracy of the clusters. A detailed description of the methods and results of the sensitivity analysis is given in Section VI.

V. Results

A. Best set of measures

The iterative process shown in Figure 8 found a set of 8 measures that best describes the expertise level of the ATCos in the dataset described in Section III (Table VI).

TABLE VI: Best set of measures found by the genetic algorithm.

M1	Ratio between the number of given DCT commands and the total number of given DCT, HDG, EFL and SPD commands
M2	Ratio between the number of given EFL commands and the total number of given DCT, HDG, EFL and SPD commands
M3	Ratio between the number of given HDG commands and the total number of given DCT, HDG, EFL and SPD commands
M4	Ratio between the total sum of squared track miles when a level command is given and the total sum of squared track miles when a level, heading or speed command is given
M5	Ratio between the total sum of squared track miles when a heading command is given and the total sum of squared track miles when a level, heading or speed command is given
M6	The aircraft with the highest flight level of all aircraft flying to waypoint MIFA
M7	The mean over all logpoints of the average TCPA per logpoint
M8	The maximum over all logpoints of the average TLOS per logpoint

Figure 9 shows the boxplots of the actual data of the individual measures from Table VI. In the figure it can be seen that there is a difference between two groups for the measures *M1*, *M2*, *M4*, *M5* and *M7*. In *M2*, *M4* and *M5* this is a difference between the expertise groups. In *M1*, *M6* and *M7* this is a difference between a mix of expertise. In *M3* and *M8* there is no clear distinction between two groups.

When looking at the boxplots of the actual data of the other measures (described in Section IV) that are not part of the best set, there was no distinct difference between two groups of ATCos for the majority of the measures. Therefore, there is no difference between expertise groups observed for these measures. Two measures are an exception: there was a difference observed between two groups in the total number of given commands and the total track penalty when using heading, speed of level commands. These two measure were not chosen by the genetic algorithm to be part of the best set of measures.

When clustering the complete set (training and test set) using the 8 measures from Table VI a heatmap with dendrogram can be created (Figure 10). From the dendrogram on top of the heatmap it can be seen that two clusters are formed: one cluster with only pro ATCos (the pro-cluster) and one cluster with mainly course ATCos (the course-cluster). While the accuracy of the pro-cluster is perfect, the accuracy of the course-cluster is not perfect because of the presence of a single pro ATCo. Furthermore, each row in the heatmap corresponds to the scaled value of the corresponding measure for each ATCo and each scenario. The dashed line in each row corresponds to the average value of the course-cluster.

Another few observations can be made from Figure 10. The most right scenario in the pro-cluster, a scenario of *P4*, has a relatively large distance between itself and the rest of the scenarios in the cluster. In the course-cluster, all scenarios of *C3* are close to each other, but are in its entirety far away from the rest of de scenarios in the course-cluster.

Each single colored rectangle corresponds to the value of the individual measure of each individual ATCo and scenario. The color is red when the measure is below course-cluster average and blue when the measure is above course-cluster average. Furthermore, areas of equally colored rectangles can be observed per measure. This indicates that, per measure, equally colored ATCos show the same behavior. The black line in each measure row shows the size of the standardized value relative to the other ATCos and scenarios in the same measure row.

It is expected that the measure results of the ATCos in the course cluster will be close to the dashed line. In general, this can be observed in the heatmap for all measures in the course-cluster. Furthermore, it is expected that the ATCos in the pro-cluster will differ uniformly from the dashed line. This can clearly be observed in measure $M2$, $M4$, $M5$ and $M1$. In the other measures this is not observed and the results are either close to the dashed line (like in measure $M8$ and $M6$) or are not uniformly distant from the dashed line (like in measure $M7$ and $M3$)

B. Differences per measure

In the boxplot of *M1* it can be seen that pro-ATCos use relatively less DCT commands compared to the course-ATCos. This difference is caused by a difference in strategy and training [6]. Pro-ATCos do not want to increase their workload by giving DCT commands while the aircraft is already flying roughly in the correct direction [6]. Course-ATCos tend to handle all aircraft more perfectly and aimed the aircraft exactly to its destination using DCT commands [6].

Looking at the boxplot of *M2* it can be seen that pro-ATCos use relatively more EFL commands compared to the course-ATCos. There was no distinct difference between the amount of given EFL commands across the expertise groups [6]. Therefore, this is linked to the fact that pro-ATCos use relatively less other command types (DCT, HDG or SPD commands).

The boxplot of M4 shows that pro-ATCos have a relative higher track penalty when using level commands compared to course-ATCos. This can be linked to *M2*. Because pro-ATCos use relative more EFL commands it can also be expected that the relative track penalty for using level commands is higher compared to course-ATCos. The same explanation can be given for the boxplot of M5. The heading commands exist of DCT and HDG commands. Because there was no distinct difference between the given HDG commands across the expertise groups [6], the results from *M5* mostly depend on the DCT commands. Since course-ATCos use relative more DCT commands it can also be expected that the relative track penalty for using heading commands is higher compared to pro-ATCos.

There is an indication that the *M1*, *M2*, *M4*, *M5* are correlated. Figure 11 shows the correlation matrix of *M1*, *M2*, *M4*, *M5*, and shows that a correlation exists between the measures. Although high correlated measures ($|\rho| > 0.99$) were removed before the selection and clustering, correlated measures still exist in the outcome of the measure selection. To see if these correlated measures only amplify other measures or will actually contribute to the difference between all ATCos, the differences across measures need to be checked.

C. Differences across measures

Although *M1* and *M5* have a high positive correlation ($|\rho| = 0.8$) differences can be seen across these measures. Looking at ATCo *C3* in *M1*, it can be seen that this ATCo differs from the rest of the course-group and behaves more like the ATCos in the pro-group. But looking at ATCo *C3* in *M5*, it can be seen that this ATCo behaves more like the rest of his or her course-group. Furthermore, this behavior is also visible for ATCo *P2*: in *M1* the ATCo behaves like the rest of his or her pro-group, but in *M5* he or she is close to the course-group. This means that although the correlation is large ($|\rho| = 0.8$) one of the measures cannot be excluded, because information about the individual ATCos will be lost.

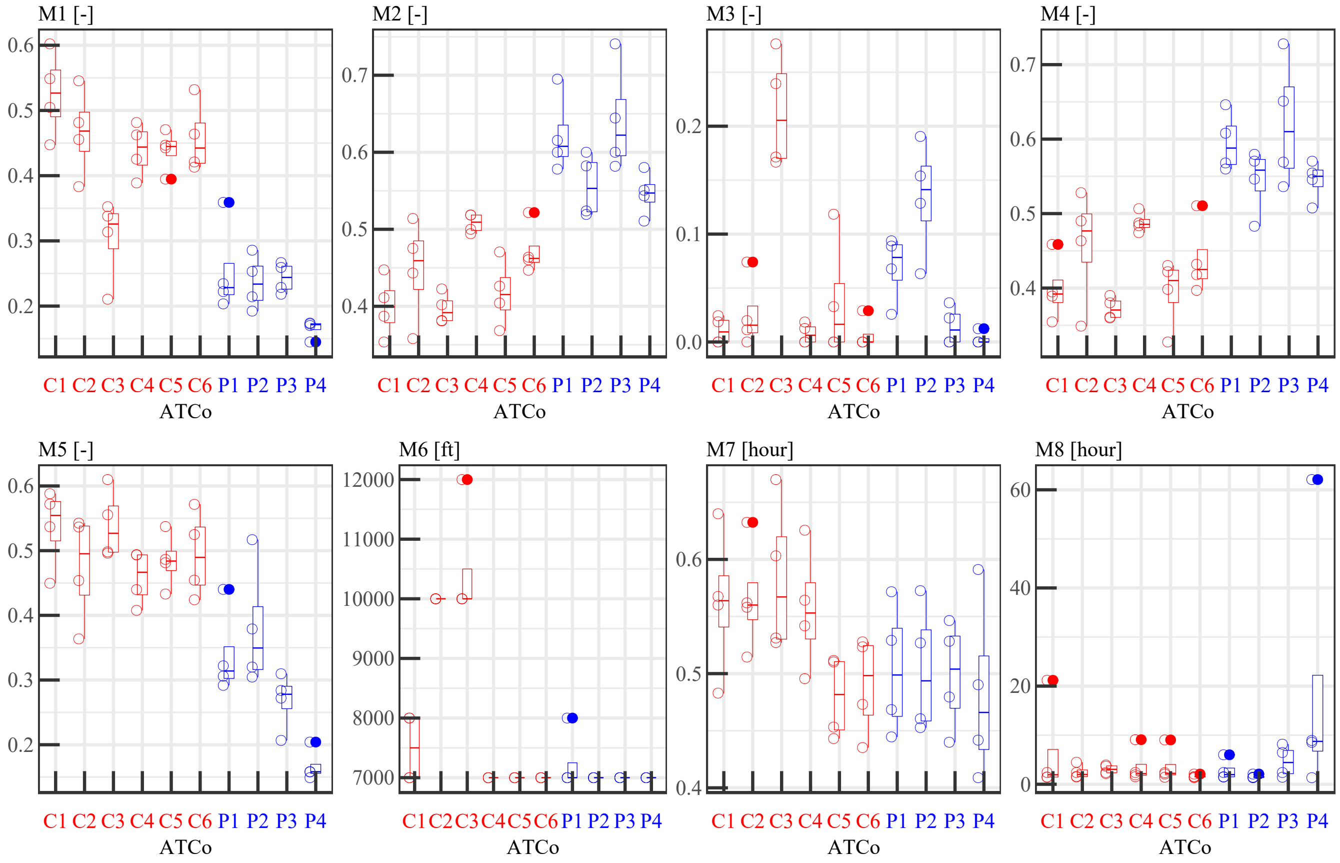


Fig. 9: Boxplots showing the individual results of the measures from the best set.

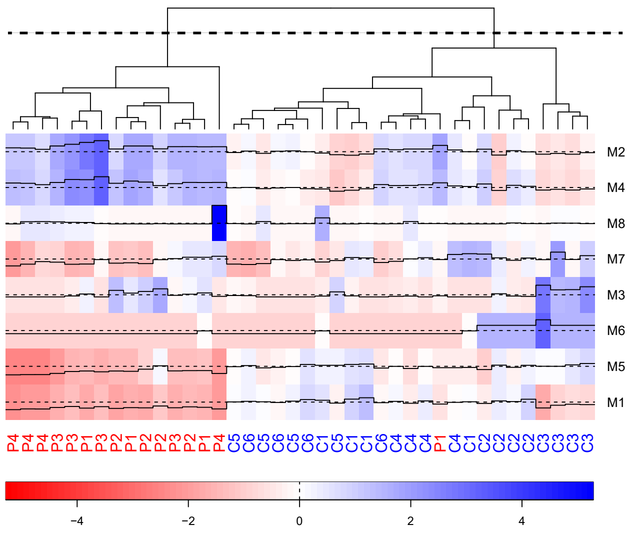


Fig. 10: Heatmap and dendrogram containing the clustering results and individual measure results.

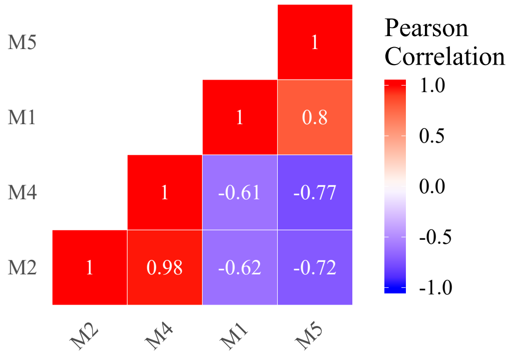


Fig. 11: The correlation matrix showing the correlation between the measures *M1*, *M2*, *M3* and *M4*.

The correlation between *M2* and *M4* is even higher ($|\rho| = 0.98$) compared to the correlation between *M1* and *M5* ($|\rho| = 0.8$). Looking at the differences across those measures, there are no distinct differences between the ATCos. Since the track penalty when using level commands depends only on the given EFL commands, *M4* is close related to the ratio of given EFL commands (*M2*). This is different compared to *M5* where the track penalty does not only depend on the given DCT commands but also on the given HDG commands. Looking at the boxplot of the ratio of given HDG commands (*M3* in Figure 9) it can be seen that the ratios of *C3* and *P2* differs significantly from its corresponding expertise group. This results in the difference between *C3* and *P2* across M1 and *M5*. Since the correlation between *M2* and *M4* is high ($|\rho| = 0.98$) and the difference across the measures is not significant, one of the measures could be excluded.

VI. Sensitivity Analysis

Since hierarchical clustering is a hard-clustering method every ATCo and scenario is always assigned to a single cluster. Even when there is no relationship in the measures, every ATCo and scenario is still assigned to a cluster. This can result into clusters that do not represent the actual relationship in the data which are called unstable clusters. Therefore, the stability of the clusters needs to be assessed. From Figure 9 it can be seen that there is a difference in how large the difference is between the two ATCo groups. Therefore, it is assessed what the contribution of the measures are to the clustering results. Furthermore, from Figure 9 it can be seen that there are (small) differences between all ATCos within each measure. Therefore, the contribution of the ATCos to the clustering results are assessed.

A. *Stability of clusters*

Since the stability of the clusters is part of the fitness function of the genetic algorithm, maximum stability is already accounted for during the selection of measures. However, this was only performed on the training set and not the complete set. Non-parametric bootstrapping (as described in Section IV) is performed on the complete set to get the stability of the clusters. Table VII shows the results from the bootstrapping process using 1000 runs.

TABLE VII: Cluster bootstrapping results.

	Cluster 1	Cluster 2
Stability	0.94	0.91

Looking at the stability of both clusters, both have a value above 0.85. This means that ATCos within each cluster highly resemble each other and show comparative behavior.

B. *Contribution of measures*

The accuracy of the clusters of all possible measures, ranging from a combination of 1 measure to a combination of 8 measures, is calculated. This results in 255 possible combinations of measures and each with its own clustering accuracy. Figure 12 shows the average accuracy of the clusters compared to the number of measures used in a combination.

In Figure 12 it can be seen that the average accuracy of the cluster shows a decreasing pattern when using less measures to cluster the ATCos (from a maximum of 0.975 to a minimum of 0.73). When removing more measures, valuable information about the distinction between the expertise groups will be lost. However, this decrease in average accuracy is negligible ($> 1\%$) when omitting just one measure.

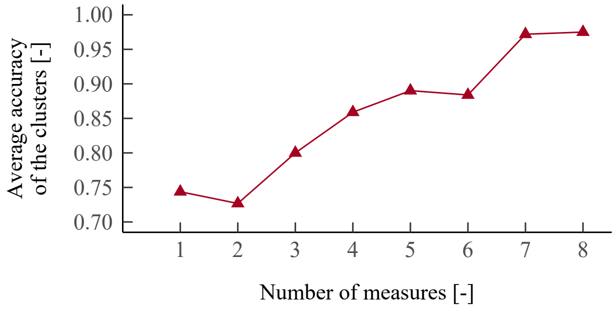


Fig. 12: Average accuracy of the clusters compared to the number of measures in a combination.

When more measures are omitted the average accuracy decreases relatively more. Then it is important to look which measures have a high contribution to the accuracy. Table VIII shows the average clustering accuracy of the 255 combinations for each measure when it is present and not present in those 255 combinations. The percentage difference between the *"With"* and *"Without"* values is shown in *"Δ%"*

TABLE VIII: Average clustering accuracy of the 255 possible measure combinations for each measure when it is present and not present in the 255 possible measure combinations. The percentage difference between the *"With"* and *"Without"* values is shown in "Δ%".

	M1	M2	M3	M4	M5	M6	M7	M8
With [-]	0.93	0.90	0.82	0.88	0.89	0.80	0.83	0.85
Without [-]	0.75	0.79	0.87	0.80	0.80	0.88	0.85	0.84
Δ%	−18.7	−12.2	6.3	−9.1	−10.2	9.9	2.3	−0.9

Looking at Table VIII it can be seen that the average accuracy drops down significantly when omitting *M1*, *M2*, *M4* and *M5*. These are the measures that have a uniform distance between the expertise groups as discussed earlier and strongly defines the differences between the two clusters. The other measures either have a positive result on the average accuracy (*M3*, *M6* and *M7*) or a small negative result on the average accuracy (*M8*) when they are omitted. Removing the measures that have a positive effect on the average accuracy when they are omitted will however not increase the average accuracy. The average accuracy drops slightly from 0.975 to 0.95 when they are all omitted.

C. *Contribution of the ATCos*

The accuracy of the clusters of all possible ATCos, ranging from a combination of 2 ATCos to a combination of 10 ATCos, is calculated. A combination of ATCos must always exist out of a minimum of one pro-ATCo and one course-ATCo. This resulted in 945 possible combinations of ATCos and each with its own clustering accuracy. Figure 13 shows the average accuracy of the clusters compared to the number of ATCos that participated in a combination. The figure also shows the number of possible combinations can be made per number of ATCos in a combination.

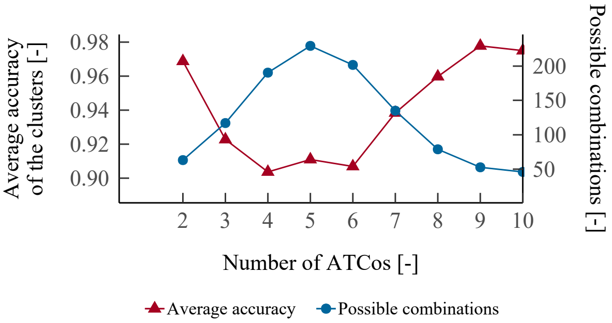


Fig. 13: Average accuracy of the clusters and the possible combinations compared to the number of ATCos in a combination.

At first, in Figure 13, it can be seen that the accuracy shows a decreasing pattern when the number of ATCos decreases. However, between 4 and 6 ATCos the accuracy stabilizes and with less than 4 ATCos the accuracy increases again. The opposite behavior is shown when looking at the possible combinations of ATCos. A high accuracy is expected when using a number of ATCos close to 10, because the selection was mainly based on the accuracy of the 10 clustered ATCos. A high accuracy is also expected when using 2 ATCos, because they will always exist of one course-ATCo and one pro-ATCo. The lower accuracies between a number of 3 and 8 ATCos are likely caused by the size of the possible combinations. In the figure the accuracy range is from 0.98 to 0.90, so the accuracy stays large across the number of ATCos. When removing just one ATCo, the accuracy changes negligibly ($> 1\%$).

Removing more ATCos will lead to a relative larger decrease in accuracy. Then it is important to look which ATCos contribute the most to the accuracy. Table IX shows the average clustering accuracy of the 945 combinations for each ATCo when it is present and not present in those 945 combinations. The percentage difference between the *"With"* and *"Without"* values is shown in "Δ%"

Looking at Table IX it can be seen that the accuracy decreases significantly when omitting *C1*, C2, *P3* and *P4*. These 4 ATCos have a large contribution to the formation of the two expertise clusters. The accuracy does not change greatly when looking at *C5*, *C6*, *P1* and *P2*. When looking at *C3* and *C4* the accuracy increases when omitting these ATCos. To explain this increase, the sum squared distances between the ATCos and the expertise clusters are compared.

TABLE IX: Average clustering accuracy of the 945 possible ATCo combinations for each measure when it is present and not present in the 945 possible ATCo combinations. The percentage difference between the *"With"* and *"Without"* values is shown in "Δ%".

	C1	C2	C3	C4	C5	C6	P1	P2	P3	P4
With [-]	0.96	0.94	0.86	0.89	0.92	0.90	0.92	0.92	0.94	0.94
Without [-]	0.88	0.89	0.98	0.95	0.92	0.94	0.91	0.91	0.89	0.89
Δ%	-8.0	-5.6	13.2	6.9	0.4	5.0	-1.4	-0.8	-6.1	-5.8

For each scenario of an ATCo, the distances of all measures to the centroid of the pro-cluster are squared and summed together. After that, the sums of squared distances are taken together to get a single sum of squared distance for each ATCo. The same is done for the distances of all measures to the centroid of the course-cluster. Table X shows the resulting sum of squared distances.

TABLE X: Sum of squared distances between the ATCos and the course-cluster centroid and the pro-cluster centroid.

	C1	C2	C3	C4	C5	C6	P1	P2	P3	P4
Pro-cluster [-]	84,0	77,0	117,2	34,6	56,3	45,7	9,4	15,0	10,7	36,4
Course-cluster [-]	15,3	19,2	49,3	11,6	12,7	11,5	50,5	40,0	67,6	92,1

Looking at *C3* in Table X it can be seen that he or she is relative distant (117,2 units) to the pro-cluster compared to the rest of the course-group. Furthermore, *C3* is also relative distant (49,3 units) to his or her own course-cluster compared to the rest of the course-group. Therefore, it can be stated that *C3* does not generally behave like a course-ATCo nor a pro-ATCo. Omitting *C3* will therefore prevent the clustering algorithm to categorize *C3* and will therefore lead to higher accuracies.

Looking at *C4* in Table X it can be seen that he or she is relative more close (34,6 units) to the pro-cluster compared to the rest of the course-group. Furthermore, *C4* is also still close (11,6 units) to the course-cluster compared to the rest of his or her course-group. Therefore, it can be stated that *C4* behaves like a course-ATCo, but also shows signs of moving toward a pro-ATCo. This is again not clear behavior for the clustering algorithm, just like *C3*. Omitting *C4* will therefore prevent the clustering algorithm to categorize *C4* and will therefore lead to higher accuracies.

VII. Discussion

To assess the capabilities of an ATCo student more frequently and at an earlier stage in training, objective measures are needed. Current objective measures are available which describe what good control behavior is.

However, it is not known which combination of objective measures could determine the expertise level of an ATCo and could be used to determine the progress of the student. This research aimed to identify a set of objective measures that can classify an air traffic controller's level of control expertise by using machine learning techniques.

Looking at the boxplots of the individual results of the 8 measures selected by the genetic algorithm (Figure 9), it can be seen that there is a preference for measures that have a distinction in the results between the expertise groups. This is desired, because these measures could be used to classify an ATCo's expertise level. The stronger these distinctions are, the more it contributes to the separation between the two expertise clusters. Furthermore, removing these preferred measures will lead to a decrease in clustering accuracy. Although the genetic algorithm found a good set of measures, by analyzing results of all the measures directly, to search for measures with a distinction in the results between the two expertise groups, is probably a quicker method without using a genetic algorithm.

Since the genetic algorithm showed a preference to measures with a distinction in the results between the expertise groups, measures *M3* and *M8* are oddly chosen. These two measures do not show a distinction between groups, but are probably selected to gain a higher fitness value for this dataset. Therefore, the results of this algorithm show signs of overfitting.

Not all results of the measures are as expected from the metrics. Looking at the ratio between the number of given DCT and EFL commands (measures *M1* and *M2*, respectively), these show a consistency in the type of instructions within each expertise group. This is not observed in the ratio between the number of given HDG commands (measure *M3*).

Measures *M4* and *M5* only show the ratio of the trackpenalty when using heading or level commands, but do not describe the moment of traffic handling. A measure that does describe this metric is the total trackpenalty in the scenario. Although this was part of the set of 55 measures, this measure was not selected by the genetic algorithm. In this measure, the difference between the expertise groups was less significant than the difference showed in the trackpenalty ratios.

Measure *M6* shows that most ATCos are generally consistent in the maximum flight level of aircraft flying to waypoint MIFA leaving the sector. However, it is more interesting to look at the consistency in the flight level instead of the maximum when looking at the variability in procedures, as discussed in Section II. Although the standard deviation of this flight level was part of the set of 55 measures, this measure was not selected by the genetic algorithm, because the difference between the expertise groups was less significant than the difference showed in the maximum of this flight level.

Measure *M7* shows a lower mean over all logpoints of the average time to closest point of approach (TCPA) per logpoint for all pro-ATCos and ATCo *C5* and *C6*. This same pattern is observed for the mean over all logpoints of the average distance at closest point of approach (DCPA) per logpoint. This could be caused by a greater experience of these ATCos. More experienced ATCos are more experienced in estimating future aircraft positions and therefore probably more willing to let aircraft pass closer to each other.

Just like measure *M3*, the maximum over all logpoints of the average TLOS per logpoint (measure *M8*) does not show a difference between the expertise groups.

The results of these 8 measures could lead to limited feedback that could be provided to the ATCos. This is not only caused by the measures that do not describe its corresponding metric, but also the correlation between measures. Only general feedback could be provided by letting the ATCo look at the use of level versus heading changes. Furthermore, the mean over all logpoints of the average TCPA per logpoint could be used as an indication for the experience gained resulting in letting aircraft pass closer to each other.

Since Ward's method aims for compact expertise clusters, there was a possibility that little variation existed within each expertise cluster. This was not observed in the results of the 8 measures as ATCo *C4* could be categorized to both expertise groups, and ATCo *C3* could neither be categorized to both expertise groups. Therefore, for this dataset, Ward's method showed desired flexibility for uncertainly placed ATCos.

However, when omitting measure *M3* it is likely that ATCo *C3* will move closer to the course-cluster, because *C3* differs significantly from the rest of the course-group in the results of measure *M3*. Since the selection of *M3* is probably caused by overfitting, the position of *C3* is therefore probably also caused by overfitting. Therefore, it cannot be stated that the genetic algorithm found a new expertise cluster with *C3* being the only ATCo in this cluster, beside the course- and pro-cluster.

Furthermore, when omitting measure *M8*, the scenario of *P4*, that is relatively distant from its pro-cluster (as shown in Figure 10), will likely move closer to the rest of the pro-cluster.

Omitting both measures *M3* and *M8* will therefore likely lead to more compact clusters, but it is not known what the impact will be on the accuracy of the clusters.

The chosen measures are highly scenario dependent, because the decisions made by the ATCos that lead to these results depend on the operational situation [1]. When adding new ATCos to the dataset it is therefore important that the new ATCos also solve the traffic problems in the same scenario when using sector dependent measures.

When comparing new ATCos (that solved the same scenarios) to the ATCos in the dataset it is not needed to cluster the data again. For comparison, the expertise-clusters are already fixed and just the boxplots of the individual measures are sufficient to give feedback to the new ATCo.

When knowing the actual expertise of the new ATCo and it needs to be appended to the dataset, then clustering is needed to determine the new clusters and the centroid of the clusters.

When adding more and more new ATCos to the dataset there is also a possibility that the current measures are not sufficient anymore. Certain specific behavior of the ATCos in this dataset could not emerge anymore and the clustering accuracy decreases too much. When this happens a different set of measures needs to be found that can describe the accuracy of this greater group of ATCos. This will be an iterative task dependent on the frequency of newly added ATCos. The advantage is that this new set of measures describes the expertise of a greater group of ATCos. When new ATCos are compared again to the greater set of ATCos his or her expertise classification will be more accurate. A greater confidence in the measures can therefore be developed as the dataset of ATCos increases.

Since analyzing the results of all the measures directly is probably quicker than using a genetic algorithm, hierarchical clustering could still be valuable. By clustering the results for the measures that have a distinction in the results between the expertise groups, the centroids of the clusters and the positions of the ATcos relative to those clusters can be determined. This could give information about the overall expertise level of an ATCo compared to the average course- or pro-ATCo.

VIII. Conclusion

In this research a method is described how machine learning can be used to find a set of measures that best describes the expertise level of an air traffic controller (ATCo). Not only a genetic algorithm is used to search for the best set of measures, but also an hierarchical agglomerative clustering algorithm (using Ward's method and a Euclidean distance measure) is used to determine the performance of that set of measures.

From prior research, a total set of objective measures was constructed which can describe good control behavior. With use of the dataset containing data of 10 ATCos (6 course-ATCos and 4 pro-ATCos) a set of 8 measures was found that can cluster the 10 ATCo's in the two expertise groups very accurately (97,5% accuracy). The genetic algorithm showed a preference for measures that have a distinction in the results between the expertise groups. Therefore, by analyzing the results of all the measures directly, to search for these type of measures, is probably a quicker method. Using hierarchical agglomerative clustering to cluster the results for these type of measures could still be used to get valuable information about the overall expertise level of an ATCo compared to the average course- or pro-ATCo.

However, not all selected measures show a difference between the expertise groups. These measures probably contribute to a better fitness, resulting in signs of overfitting. These measures, together with correlated measures, also result in limited feedback that can be provided to the ATCos.

For future research it is recommended to determine the clustering performance of the set of measures that have a distinction in the results between the expertise groups. When this performance is comparable to the best set found by the genetic algorithm, the use of a genetic algorithm could be replaced by directly selecting the measures that show this distinction.

References

[1] M. J. Schuver-van Blanken, H. Huisman, and M. I. Roerdink, "The ATC Cognitive Process and Operational Situation Model - A model for analysing cognitive complexity in ATC," in *29th EAAP Conference*, Budapest, Hungary, 2010, Conference Paper.

[2] M. J. Schuver-van Blanken and M. I. Roerdink, "Clarifying Cognitive Complexity and Controller Strategies in Disturbed Inbound Peak ATC Operations," in *17th International Symposium on Aviation Psychology*, Dayton, OH, 2013, Conference Paper.

[3] B. Flanagan, S. Hirokawa, E. Kaneko, and E. Izumi, "Classification of Speaking Proficiency Level by Machine Learning and Feature Selection," in *1st International Symposium on Emerging Technologies for Education, SETE 2016 Held in Conjunction with ICWL 2016*, vol. 10108 LNCS. Rome, Italy: Springer, 2017, Conference Paper, pp. 677–682.

[4] G. Boccignone, M. Ferraro, S. Crespi, C. Robino, and C. De'Sperati, "Detecting expert's eye using a multiple-kernel Relevance Vector Machine," *Journal of Eye Movement Research*, vol. 7, no. 2, 2014.

[5] R. A. Watson, "Use of a Machine Learning Algorithm to Classify Expertise: Analysis of Hand Motion Patterns During a Simulated Surgical Task," *Academic Medicine*, vol. 89, no. 8, pp. 1163–1167, 2014.

[6] V. L. J. Somers, "3D Solution Space-based Prediction of Air Traffic Control Workload," Unpublished M.Sc. Thesis, Faculty of Aerospace Engineering, Delft University of Technology, 2017.

[7] E. Oprins and M. Schuver, "Competentiegericht opleiden en beoordelen bij LVNL (Competence-based training and assessment at LVNL)," Human Factors Advisory Group, Schiphol, The Netherlands, Newsletter Article 6, 2003.

[8] E. Oprins, E. Burggraaff, and H. van Weerdenburg, "Design of a Competence-Based Assessment System for Air Traffic Control Training," *The International Journal of Aviation Psychology*, vol. 16, no. 3, pp. 297–320, 2006.

[9] K. Kallus, D. van Damme, and A. Dittmann, "Integrated Task and Job Analysis of Air Traffic Controllers – Phase 2: Task Analysis of En-route Controllers," EUROCONTROL, Brussels, Belgium, Technical Report, 1999.

[10] B. Hilburn, "Cognitive Complexity in Air Traffic Control: A Literature Review," EUROCONTROL Experimental Centre, Brétigny-sur-Orge, France, Technical Report, 2004.

[11] M. J. Schuver-van Blanken and J. G. van Merriënboer, "Air Traffic Controller Strategies in Operational Disturbances - An exploratory study in air traffic control," in *30th EAAP Conference*, Sardinia, Italy, 2012, Conference Paper.

[12] B. Kirwan and M. Flynn, "Identification of Air Traffic Controller Conflict Resolution Strategies for the CORA (Conflict Resolution Assistant) Project," EUROCONTROL Experimental Centre, Brétigny, France, Technical Report, 2001.

[13] ICAO, "Doc 4444, Procedures for Air Navigation Services - Air Traffic Management," ICAO, Quebec, Canada, Technical Report, 2016.

[14] J.-F. D'Arcy and P. S. Della Rocco, "Air Traffic Control Specialist Decision Making and Strategic Planning - A Field Survey," Federal Aviation Administration, Atlantic City, NJ, Technical Report, 2001.

[15] G. A. Mercado Velasco, M. Mulder, and M. M. van Paassen, "Analysis of Air Traffic Controller Workload Reduction Based on the Solution Space for the Merging Task," in *AIAA Guidance, Navigation, and Control Conference*, Toronto, Canada, 2010, Conference Paper.

[16] G. A. Mercado Velasco, M. Mulder, and M. M. van Paassen, "Air Traffic Controller Decision-Making Support using the Solution Space Diagram," *IFAC Proceedings Volumes*, vol. 43, no. 13, pp. 227–232, 2010.

[17] S. M. A. Rahman, C. Borst, M. Mulder, and M. M. Van Paassen, *Measuring Sector Complexity: Solution Space-Based Method*, ser. Advances in Air Navigation Services. InTech, 2012.

[18] G. A. Mercado Velasco, C. Borst, J. Ellerbroek, M. M. Van Paassen, and M. Mulder, "The use of intent information in conflict detection and resolution models based on dynamic velocity obstacles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2297–2302, 2015.

[19] IenM, "Luchtruimvisie - Bijlagerapport 1: Huidige inrichting en beheer van het Nederlandse luchtruim," Ministerie van Infrastructuur en Milieu, Den Haag, The Netherlands, Technical Report, 2012.

[20] S. Fothergill and A. Neal, "Conflict-resolution heuristics for en route air traffic management," in *57th Human Factors and Ergonomics Society Annual Meeting - 2013, HFES 2013*. San Diego, CA: Proceedings of the Human Factors and Ergonomics Society Annual Meeting 57, 2013, Conference Paper, pp. 71–75.

[21] R. Xu and D. Wunsch Ii, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[22] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is "Nearest Neighbor" Meaningful?" in *Database Theory — ICDT'99*, C. Beeri and P. Buneman, Eds. Heidelberg, Germany: Springer, 1999, Conference Paper, pp. 217–235.

[23] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction To Cluster Analysis*, 1st ed. New York, NY: Wiley, 1990.

[24] M. Negnevitsky, *Artificial Intelligence: A Guide to Intelligent Systems*, 3rd ed. Harlow, UK: Addison Wesley, 2011.

[25] R. Kohavi and F. Provost, *Glossary of Terms*, 1998, vol. 2.

[26] C. Hennig, "Cluster-wise assessment of cluster stability," *Computational Statistics and Data Analysis*, vol. 52, no. 1, pp. 258–271, 2007.

Part II

Preliminary Report

(Previously graded under AE4020)

Chapter 1

Introduction

1-1 Background

In order to reach a safe, orderly and expeditious flow of aircraft within a Control Area (CTA), Area Control Centers (ACCs) rely on the expertise of their Air Traffic Controllers (ATCos) (ICAO, 2016). Due to the highly complex and dynamic environment, ATCos need to process large amounts of dynamically changing information while maintaining a good balance between safety, efficiency and environment (Schuver-van Blanken et al., 2010; Oprins et al., 2006). Furthermore, human error is not allowed because of the strict safety requirements, even while the workload has increased during the last decade (Oprins et al., 2006).

The selection of candidates for the Air Traffic Control (ATC) training is strict, because of the high demanding nature of the job. Furthermore, these candidates need to acquire the required competences during the limited training period (usually two or three years). Unfortunately, even after a strict selection, a high drop-out rate is present during the training period, because the required expertise level is not reached or cannot be reached within the training period (Schuver-van Blanken & Roerdink, 2013). This is undesirable, because a large amount of time, effort and money has been put into these students.

The determination of the level of expertise is currently done by using subjective assessments. The instructor assesses the level of the students based on his or her own experience. Therefore, this assessments will always have a subjective influence. To remove this subjective influence, objective assessments are needed. These assessments consist of metrics with their corresponding objective measures. Objective assessments might result in a reliable and more fair measure of the capabilities of a student. Furthermore, objective assessments might give the ability to determine the competences that are still underdeveloped, so that more attention can be given to these competences during training (Oprins & Schuver, 2003). To find the combination of which competences, which set of metrics, and which set of objective measures that determine the levels of expertise accurately is part of an extensive search.

Conducting such an extensive search is a task that might be too extensive and complex for the human mind to solve. A solution to this problem is using machine learning algorithms.

Machine learning algorithms are capable of finding patterns, classifying objects, or learning specific behavior by getting rewards or punishments (Russell & Norvig, 2010). The purpose for which machine learning techniques are intended give good prospects for the determination of which set of competences, metrics and objective measures determine the expertise levels.

1-2 Problem Description

It is currently known that certain competences can be objectively measured, but not which particular set of objective measures might determine whether an ATCo is a novice, an intermediate or an expert. Since there are many objective measures in ATC, a good set takes a considerable amount of time to find.

Machine learning could be a solution to get good and quick results when one searches for a good set of objective measures. Studies that use machine learning in order to determine experience level have already been performed on speaking proficiency levels (Flanagan et al., 2017), billiard players (Boccignone et al., 2014) and surgeons (Watson, 2014). All with promising results. This gives a good outlook that machine learning techniques might also be applied to identify a set of measures that is able to determine an ATCo's level of control expertise. The focus in this thesis will be on the cluster analysis techniques of machine learning.

The goal is that the machine learning algorithm must be able to classify the data into the expertise groups using a specific set of objective measures. The research objective of this thesis is to identify a set of objective measures that can classify an air traffic controller's level of control expertise by using a cluster analysis machine learning technique.

1-3 Research Questions

Following the research objective, the research in this thesis is centered around a main research question:

> *How can a clustering algorithm be used to determine a set of objective measures, based on good control behavior, that accurately describe an air traffic controller's expertise level?*

In order to answer the main research question, five subquestions are defined:

1. (a) Which objective measures can accurately describe good control behavior?
   (b) Which clustering algorithms can be used to determine a set of objective measures?
2. Which clustering algorithm has the best performance in clustering the data into different expertise groups?
3. What is the performance of the clustering algorithm when it is subjected to unseen data?
4. Does the set of objective measures accurately describe an air traffic controller's expertise level?

1-4 Research Approach

To answer the main research question and subquestions first a literature review is conducted. In this literature review it is analyzed what good control behavior looks like and what objective measures can be used to describe this behavior. Furthermore, the available clustering algorithms are analyzed and explained. The advantages and disadvantages of each algorithm will be discussed, and the most promising algorithm is chosen.

Next, data preparation is conducted on the available dataset, containing the data of the professional group and the ATC course/research group. Based on the objective measures from the literature review, the appropriate data is selected. This data is then preprocessed in order to format the data into a workable form. As a final step in the data preparation, the preprocessed data is transformed in order to use this data as an input in the clustering algorithm. By testing the selected algorithms from the literature review, with the prepared data, it is determined which algorithm and set of objective measures can perhaps cluster the data into the two expertise groups.

After the testing of the clustering algorithm, an experiment is performed on novices. They will solve the same traffic scenarios as the professional group and the ATC course/research group. This generated dataset is used, after data preparation, to determine the performance of the clustering algorithm with the set of objective measures when it is subjected to unseen data.

Finally, conclusions are drawn to determine whether the clustering algorithm can cluster all the prepared data into three expertise groups. An answer can be given to the question if the set of determined measures can accurately describe an ATCo's expertise level.

To limit the scope of this research project, several assumptions are made. First, the focus is solely on objective measures. These measures can be directly gathered from the data or constructed from the available data. Furthermore, the values of these measures are not directly provided by the participant to remove the subjective influence. Second, a limited set of measures and a limited number of samples are available in the dataset. Therefore, the selection of the measures is based on the availability of the data in the dataset.

1-5 Report Structure

This report contains the preliminary part of this thesis. The preliminary phase consists of a literature review and a preliminary analysis. Chapter 2 contains the literature review about ATCo behavior and the corresponding objective measures. Chapter 3 contains the literature review about available clustering algorithms. Chapter 4 contains an analysis of the selected clustering algorithm, objective measures and the available dataset. Part of the dataset is clustered to give a practical insight in the theory from the literature. Chapter 5 contains the conclusion of the preliminary analysis, and some future steps for the main phase of this thesis.

Chapter 2

Air Traffic Controller Competences and Assessment

In this chapter the ATCo competences, ATC structural elements, and the corresponding metrics and measures are discussed. A competence is the ability to apply a combination of acquired skills, knowledge and attitudes to perform a given task (Oprins et al., 2006). This combination of skills, knowledge and attitudes must be acquired during the ATCo training. ATC structural elements are elements from the operational situation that have an influence on the competences of an ATCo.

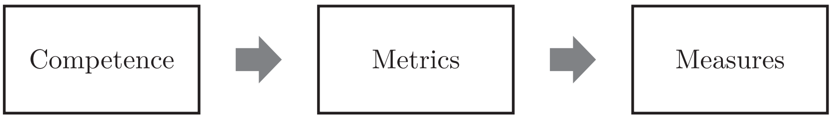


Figure 2-1: Finding objective measures for each competence based on the metrics

Figure 2-1 shows how the competences can be assessed using metrics and corresponding objective measures. For each competence there are one or more metrics available and for each metric there are one ore more objective measures available. The outcome of the objective measures are influenced by the ATC structural elements in the operational situation.

Section 2-1 describes what the goals are of an ATCo. The ATCo competences and ATC structural elements are discussed in Section 2-2 with use of the ATC Cognitive Process & Operational Situation model (ACoPOS) model (Schuver-van Blanken et al., 2010). In Section 2-3, metrics are linked to each competence and structural element. Finally, in Section 2-4, measures are linked to each metric. This finishes a list with ATCo competences and ATC structural elements, with the corresponding metrics and measures.

2-1 The Air Traffic Controller

The main goal of an ATCo is to ensure a safe, orderly and expeditious flow of air traffic in his or her sector (Oprins et al., 2006). By providing the pilots with heading, speed and altitude commands, the ATCo prevent violation of separation minimums and possible collisions. Beside the main goal, ATCos have several side goals. ATCos must also guide aircraft in an efficient manner to avoid unnecessary delays of flights. Furthermore, an ATCo must also manage his or her own mental workload, because of the high demanding nature of the job, and deal with personal and environmental influences which can have an influence on the performance (Oprins et al., 2006). These goals are reflected in the performed actions by the ATCos.

To what extent these goals are reached is determined by the expertise level of the ATCo. The level of expertise is among other things determined by experience, as well as having a natural talent to the profession. The expertise level is determined by a set of ATC related competences. High requirements are set for these competences because of the cognitive complexity for the ATCo. This cognitive complexity cannot be seen separately from the operational situation (Schuver-van Blanken et al., 2010). Therefore, both the ATCo competences and the ATC structural elements in the operational situation are discussed in Section 2-2.

2-2 ATCo Competences and ATC Structural Elements

In this section, the ATCo related competences and the ATC structural elements in the operational situation are discussed. The relationship between the competences of the ATCo and the ATC structural elements can be viewed in the ACoPOS model (Figure 2-2). In Figure 2-2, the *air traffic controller* block contains the ATCo competences and the *operational situation* block contains the ATC structural elements. The selection of the competences and the ATC structural elements will be based on the competences and the ATC structural elements present in this model.

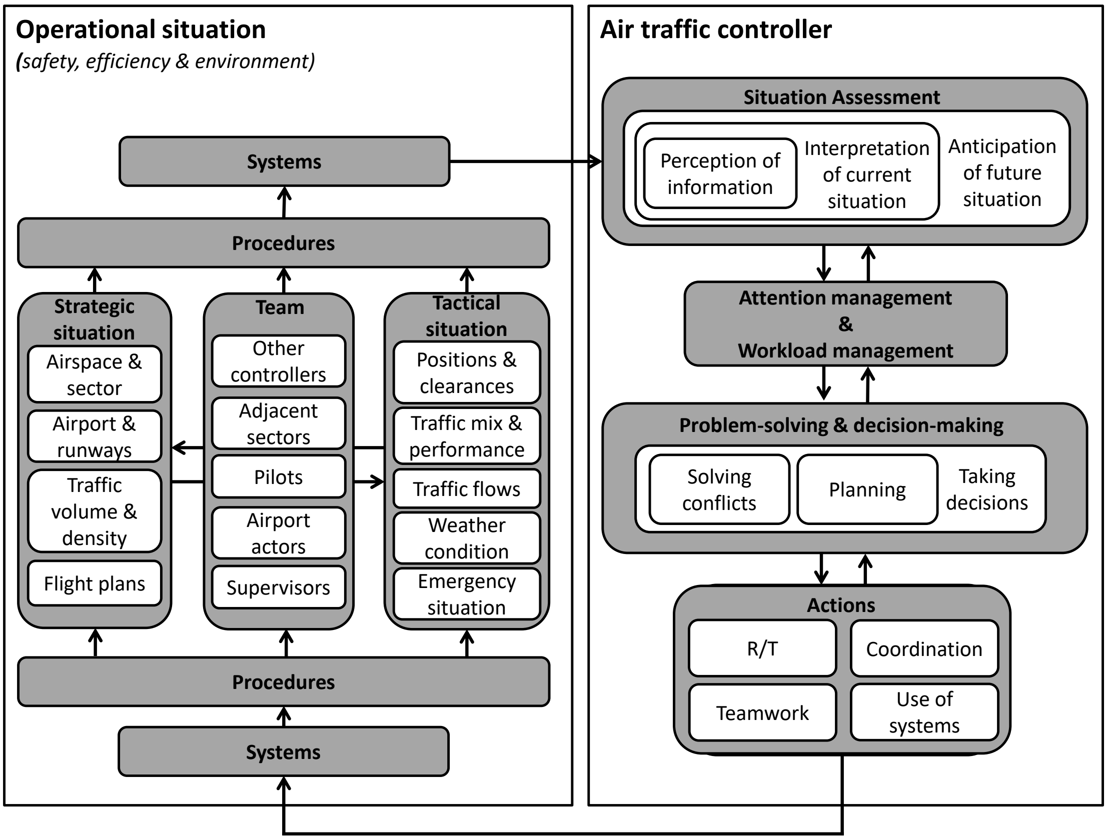


Figure 2-2: The ATCo Cognitive Process & Operational Situation model (ACoPOS model) (adapted from Schuver-van Blanken et al. 2010)

In this figure, the blocks *situation assessment*, *problem solving & decision making*, and *attention management & workload management* form a representation of the cognitive processes of the ATCo. These cognitive processes form the basis for the performed actions of the ATCo (Oprins et al., 2006). The competences present in the cognitive processes can only be assessed subjectively, but the result of the cognitive process is reflected in the performed actions, which can be objectively assessed (Oprins et al., 2006). Therefore, the competences present in the cognitive processes will not be assessed in this research.

An exception is made for the *solving conflicts* competence. This competence reflects the usage of conflict resolution strategies. According to Loft et al., a strategy is: *"a specific class of air traffic management that achieves one or more objectives (e.g., safety, orderliness, expeditiousness) with a certain investment of time and effort."* (Loft et al., 2007, p.380). The performed actions can be objectively measured to demonstrate whether certain strategies have been used.

In the ACoPOS model, the competences in the *actions* category can be objectively assessed. Radio Telephony (R/T) focuses on the interaction between the ATCo and the aircraft, while the other competences focus on interaction between ATCos or interaction between the ATCo and the equipment (Oprins & Schuver, 2003). The focus in this research will be on the interaction between the ATCo and the aircraft and therefore only R/T will be assessed in this research.

In the *operational situation* block of Figure 2-2, the ATC structural elements are present. These elements have influence on the cognitive complexity of the ATCo.

The operational situation is marked by the safety, efficiency and environment requirements. Finding a correct balance between these aspects is the core task of the ATCo (Schuver-van Blanken et al., 2010).Safety and efficiency is determined by the actions of the ATCo. To what extent safety and efficiency are reached can be determined by the flight movements and R/T recordings. Therefore, safety and efficiency are highly linked to the ATCo competences.

How the traffic needs to be handled formally inside the sector is determined by the procedures (Schuver-van Blanken et al., 2010). These procedures can limit the amount of available solutions for an ATCo. Therefore, the element *procedures* has an influence on the ATCo.

In the ACoPOS model, the strategic traffic situation determines the physical boundaries in which an ATCo needs to handle traffic (Schuver-van Blanken et al., 2010). These boundaries are determined by *traffic volume & density.* Furthermore, *airport & runway*, *flight plans*, and *airspace & sector* may limit the amount of available solutions for the ATCo. Although these last three mentioned competences could have an influence on the ATCo, no metrics or measures will be searched for these competences because of the limited availability of the data for further analysis.

In the ACoPOS model, the tactical traffic situation is marked by the dynamic and changing nature of the situation. This aspect has an effect on the cognitive complexity for the ATCo, since the results of the changing situations are often non-preferable or unexpected. These include deteriorated weather conditions and emergency situations (Schuver-van Blanken et al., 2010). Although these two mentioned elements could have an influence on the ATCo's performance, no metrics or measures will be searched for these elements because of the limited availability of the data for further analysis. *Position & clearances*, *traffic mix & performance*, and *traffic flows* also have an influence on the decision making of the ATCo and can be objectively measured.

Table 2-1 gives an overview of four ATCo competences that are assessed in this research. Furthermore, the table shows the five ATC structural elements that have an influence to what extent these competences are reached. The competences and structural elements can be objectively assessed using the metrics described in the next section (Section 2-3).

Table 2-1: Selected ATCo competences and ATC structural elements from the ACoPOS model

ATCo competences	**ATC structural elements**
• Solving conflicts • R/T • Safety • Efficiency	• Traffic volume & density • Position & clearances • Traffic mix & performance • Traffic flows • Procedures

2-3 Metrics

In order to determine to what extent the ATCo competences are reflected, metrics are linked to each competence. Each competence can have multiple metrics, because the same competence can be assessed in multiple ways. First, a selection is made which metrics can be used to assess the competences from Table 2-1. After the selection of the metrics for competences, a selection of the metrics for the ATC structural elements from Table 2-1 is made.

2-3-1 ATCo Competences

To assess the *solving conflict* competence it must be determined what conflict resolution strategies are used. The strategies described in this subsection are considered best practice in conflict resolution. Fothergill & Neal performed a study observing a series of conflict resolution strategies (Fothergill & Neal, 2013). These strategies consisted of lateral and vertical conflict resolution strategies, observed from experienced en-route ATC operators. The choice between level changes (vertical resolution) and vector/speed changes (lateral resolution) depended on the experienced workload. Level changes are more preferred when the workload is high (Fothergill & Neal, 2013). Furthermore, another best practice in conflict resolution strategies is to minimize the number of aircraft to move (Kirwan & Flynn, 2001).

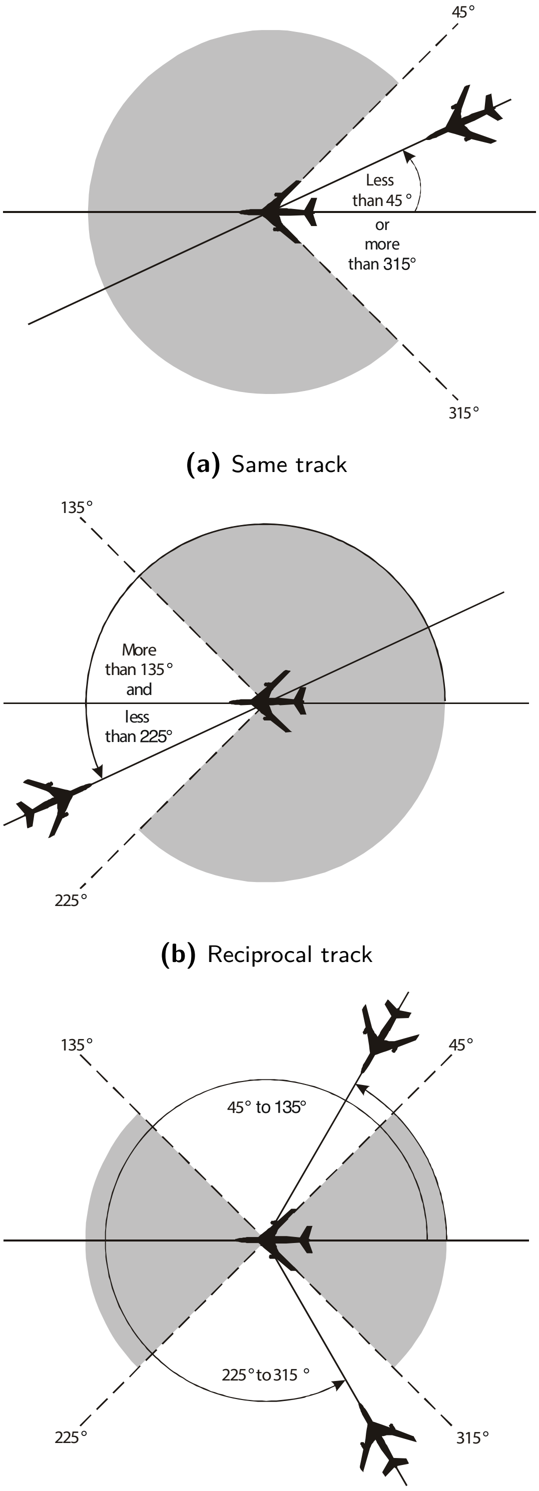
Within the vertical conflict resolution strategies, Fothergill & Neal identified two *"cut off"* strategies dependent on the available time an ATCo has (Fothergill & Neal, 2013). With less time available, ATCos were more likely to cut off at the level closest to the current level of the aircraft, compared to a cut off at the highest possible flight level on climb (Fothergill & Neal, 2013).

Other vertical conflict resolution strategies, as found by Fothergill & Neal, include: letting the aircraft descent to the nearest available level, or to ensure that every aircraft in his or her sector is at a different flight level (Fothergill & Neal, 2013). Letting an aircraft descent to a lower level will lead to a quick separation, but it is not beneficial for the fuel consumption (Fothergill & Neal, 2013). Ensuring that all aircraft are on a different level may be beneficial during high traffic load, but it is possible that an available level is not near the current level of the aircraft (Fothergill & Neal, 2013). There is a possibility that a difference exists in use of these two strategies within different expertise levels.

Under high workload, using vertical conflict resolution strategies are more preferred compared to lateral conflict resolution strategies (Fothergill & Neal, 2013). However, under extremely high workload, ATCos prefer to use lateral conflict resolution strategies (Fothergill & Neal, 2013). For example, when aircraft are in conflict, and can only climb and not descend, using lateral conflict resolution strategies give separation assurance the quickest (Fothergill & Neal, 2013).

First, to explain the lateral conflict resolution strategies, three types of conflict geometries should be distinguished. These types are described by International Civil Aviation Organization (ICAO) and are shown in Figure 2-3 (ICAO, 2016). The type of conflict geometry has an influence in the conflict resolution choice of the ATCo. The three types of conflicts include:

- Same track (heading difference which is less than 45 degrees or more than 315 degrees)
- Reciprocal tracks (heading difference which is more than 135 degrees and less than 225 degrees)
- Crossing tracks (heading difference between 45 degrees and 135 degrees, or between 225 degrees and 315 degrees)



(a) Same track

(b) Reciprocal track

(c) Crossing track

Figure 2-3: Three types of conflict geometries defined by ICAO (ICAO, 2016)

According to Kirwan & Flynn, for a same track conflict, the best practice for an ATCo is to turn the faster aircraft direct to route in front of the slower aircraft (Kirwan & Flynn, 2001). Next, for crossing tracks conflict, the best practice is to turn the slower aircraft behind the faster aircraft. Kirwan & Flynn did not define a best practice solution in case of reciprocal tracks conflicts, but only that reciprocal tracks conflicts need to be solved first of all present conflicts (Kirwan & Flynn, 2001). The lateral conflict resolution strategies found by Fothergill & Neal only describe the actual resolutions to the general principles described by Kirwan & Flynn. For example, directing a slower aircraft on a track parallel to its original track could be compared to the best practice to turn the slower aircraft behind the faster aircraft (Fothergill & Neal, 2013). In this research, only the general principles regarding the lateral strategies are assessed, since it is expected that these principles can be identified better from the data.

The difference in usage of these conflict resolution strategies may reflect the experience level of the ATCo since expert controllers are more consistent and use similar solutions compared to novices (Kallus et al., 1999).

An experienced ATCo must keep his or her workload as low as possible (Hilburn, 2004). The way R/T is used, has an influence on this workload, because communicating more with the aircraft takes extra time. Therefore, the way R/T is used to keep the workload low can be used as a metric to determine experience. According to Kallus et al., ATCos have an internal *"conflict solution library"* (Kallus et al., 1999, p.46). The most frequent and commonly used solutions come first in mind. These solutions need certain types of instructions. It is therefore expected that experienced ATCos are more consistent in the use of certain types of instructions. Finally, it is expected as an ATCo is more experienced, he or she will make less errors in the communication with aircraft. This can also be used as a metric for experience.

As stated in Section 2-1, ensuring a safe flow of air traffic is part of the main goal of an ATCo. The safety inside the sector is ensured by maintaining the separation minimums. These separation minimums are set by ICAO and described as 5NM horizontal separation and 1000ft vertical separation (ICAO, 2016). This creates a Protected Zone (PZ) around the aircraft that no other aircraft should access. Another way to ensure safety is to be more conservative or cautious, dependent on the ATCo's age and fatigue, the experienced workload, or factors like bad weather (D'Arcy & Della Rocco, 2001). Furthermore, formal operating procedures must be used inside a sector. The number of procedures, the complexity of procedures, and the diversity in working methods all have an influence on the cognitive complexity (Schuver-van Blanken et al., 2010). Therefore, with high cognitive complexity, errors in the use of procedures could emerge. This leads to the use of sufficient safety buffers when needed.

Another part of the main goal of an ATCo is to provide and efficient flow of air traffic, as stated in Section 2-1. By interviewing ATCos, Kirwan & Flynn found many principles and strategies used by ATCos (Kirwan & Flynn, 2001). One of those principles is to minimize the additional track miles flown, that reflect the efficiency competence. A metric that is related to the minimization of the additional track miles flown is to minimize the delay time of the aircraft (Oprins et al., 2006). When an aircraft needs to fly additional track miles, it is possible that a delay will occur, unless the ATCo allows the aircraft to fly faster. Finally, part of the task of an ATCo is to create an expeditious flow of air traffic in his or her sector (ICAO, 2016). A higher outflow of aircraft might indicate a higher efficiency.

The metrics that describe the ATCo competences from Table 2-1 are shown in Table 2-2.

Table 2-2: Metrics to assess ATCo competences

	Solving conflicts	
1	Level changes vs vector/speed changes	(Fothergill & Neal, 2013)
2	Minimize the number of aircraft to move	(Fothergill & Neal, 2013)
3	Cut off at nearest available level on climb	(Fothergill & Neal, 2013)
4	Cut off at highest possible level on climb	(Fothergill & Neal, 2013)
5	Descend to nearest available level	(Fothergill & Neal, 2013)
6	Assign the only level available	(Fothergill & Neal, 2013)
7	For crossing conflicts, turn slower aircraft behind	(Kirwan & Flynn, 2001)
8	For same track conflict, turn faster aircraft direct to route in front of slower aircraft	(Kirwan & Flynn, 2001)
	R/T	
9	The way of R/T use to keep the workload low	(Hilburn, 2004)
10	Consistency in the type of instructions	(Kallus et al., 1999)
	Safety	
11	Maintain separation minimums	(ICAO, 2016)
12	Use sufficient safety buffers	(D'Arcy & Della Rocco, 2001)
13	Errors in using procedures	(Schuver-van Blanken & Merriënboer, 2012)
	Efficiency	
14	Minimize additional track miles flown	(Kirwan & Flynn, 2001)
15	Minimize delay time	(Oprins et al., 2006)
16	Create an expeditious flow of traffic	(ICAO, 2016)

Beside the metrics described in Table 2-2, which are supported by literature, there are also metrics that are supported by certain observations or expectations which cannot be directly found in literature. For example, insights can be drawn from visualization of the available data. By using this method metrics are formulated based on the observations from the data. Metrics from expectations are formed based on a common sense. For example, an experienced ATCo makes less R/T mistakes compared to a novice ATCo since the experienced ATCo had more time to practice with R/T and therefore becoming better and making less mistakes.

Since an experienced ATCo has more controller experience in the field of ATC compared to a novice ATCo, it is reasonable to think that the experienced ATCo has a quicker overview of the situation and handles traffic quicker. Therefore, it is reasonable to think that an experienced ATCo will give a level, heading or speed change far before the aircraft leaves the sector. The moment of traffic handling can therefore be seen as a metric.

Table 2-3 shows metrics that are based on hypotheses instead of literature support. This table is not a complete list of metrics, because more metrics can be formed when visually inspecting the data in further analysis.

Table 2-3: Metrics from hypotheses to assess ATCo competences

	R/T
1	Minimize the amount of errors in R/T
	Efficiency
2	Moment of traffic handling

2-3-2 ATC Structural Elements

To determine to what extent the ATC structural elements from Section 2-2 have an influence on the reflection of the competences of the ATCo, metrics are linked to each element. Each ATC element can have multiple metrics, because the same element can be assessed in multiple ways. In the remainder of this section, a selection is made which metrics can be used to assess the ATC structural elements from Table 2-1

To determine the effect of the traffic volume & density on the ATCo, the availability of the solution space and the air traffic density needs to be determined. According to the findings of Schuver-van Blanken & Roerdink, ATCos create solution space or use the solution space that is already available (Schuver-van Blanken & Roerdink, 2013). The availability of the solution space has influence on the ability to use certain conflict resolution strategies, like the lateral resolutions, the efficiency and the prevention of future problems (Schuver-van Blanken & Roerdink, 2013). Furthermore, the air traffic density has an influence on the complexity of the ATCo's task and has therefore an influence on the choice of action (Hilburn, 2004).

The detection of a conflict and the choice of an appropriate resolution strategy is dependent on the position and the movement of the aircraft. An important factor is the Closest Point of Approach (CPA) and especially the Distance at Closest Point of Approach (DCPA) and the Time to Closest Point of Approach (TCPA). An ATCo uses the DCPA to determine if separation minimums are not breached. Furthermore, the TCPA has an influence in the usage of conflict resolution strategies (Fothergill & Neal, 2013). Furthermore, the conflict geometry and speed difference have an influence in conflict resolutions described by Kirwan & Flynn (Table 2-2). The ATCos in the research of Fothergill & Neal also stated that the amount of aircraft that change their flight level has an influence on the decision making (Fothergill & Neal, 2013).

Since the traffic in the sector consist of multiple types of aircraft, which do not all have the same performance in terms of speed and climb performance, the type of aircraft has an influence in the decision making of the ATCo (Fothergill & Neal, 2013). Furthermore, studies by Schuver-van Blanken & Roerdink showed that ATCos create traffic patterns or use existing patterns (Schuver-van Blanken & Roerdink, 2013). This helps them to create an overview and manage expectations (Schuver-van Blanken & Roerdink, 2013).

Within a sector, a procedure can result in several options for the ATCo. For example, an ATCo can have the option the let the aircraft leave the sector between flight level 70 and 100. This means that there is a variability in this procedure. An ATCo can decide to let all aircraft leave the sector at flight level 70, but can also let the aircraft leave the sector at different flight levels between flight level 70 and 100.

The metrics that describe the ATC structural elements from Table 2-1 are shown in Table 2-4.

Table 2-4: Metrics to assess ATC structural elements

	Traffic volume & density	
1	Availability of solution space	(Schuver-van Blanken & Roerdink, 2013)
2	Air traffic density	(Hilburn, 2004)
	Position & clearances	
3	Closest point of approach	(Fothergill & Neal, 2013)
4	Conflict geometry	(Kirwan & Flynn, 2001)
5	Speed difference	(Kirwan & Flynn, 2001)
6	Amount of aircraft changing levels	(Fothergill & Neal, 2013)
	Traffic mix & performance	
7	Type of aircraft	(Fothergill & Neal, 2013)
	Traffic flows	
8	Use of current traffic patterns	(Schuver-van Blanken & Roerdink, 2013)
	Procedures	
9	Variability in procedures	(Schuver-van Blanken et al., 2010)

2-4 Measures

From the metrics that assesses the ATCo competences, described in Section 2-3, objective measures can be linked to each metric. Measures are the values from the data, either directly obtained from the dataset or constructed from the directly available data. Objective measures are obtained from measuring equipment or ATC equipment (i.e. radar). The measures are not subjected to personal opinion or interpretation during measurement. Therefore, the objective measures could explain how well the task is performed, regardless of what the experience was while performing the task. Furthermore, it must be noted that the ATC structural elements have an influence on the decision making of an ATCo. The result of this decision making could be seen in the objective measures.

The objective measures are used in different ways to assess a specific metric. Looking at the metrics that assess the *solving conflicts* competence (Table 2-2), it mainly consists of conflict resolution strategies. To determine if a particular strategy is used, a combination of objective measurements must be analyzed. This can be seen for example, in the metric for crossing conflicts. In this metric the slower aircraft is turned behind the faster aircraft. To determine if this metric is used, not only the aircraft heading must be gathered, but also the speed of the aircraft. Furthermore, this strategy is used in a conflict situation, and therefore the position and the altitude of the aircraft must also be gathered. This shows that the performance of a metric cannot always be determined by a single measure.

Besides that a combination of multiple measures can only describe the performance of a single metric, it is also possible that individual single measures describe the performance of a metric. This can be seen in the metric that an ATCo needs to maintain separation minimums. To determine the performance of this metric, not only the number of conflicts can be measured, but also the number of mid-air collisions or the number of Loss of Separation (LOS).

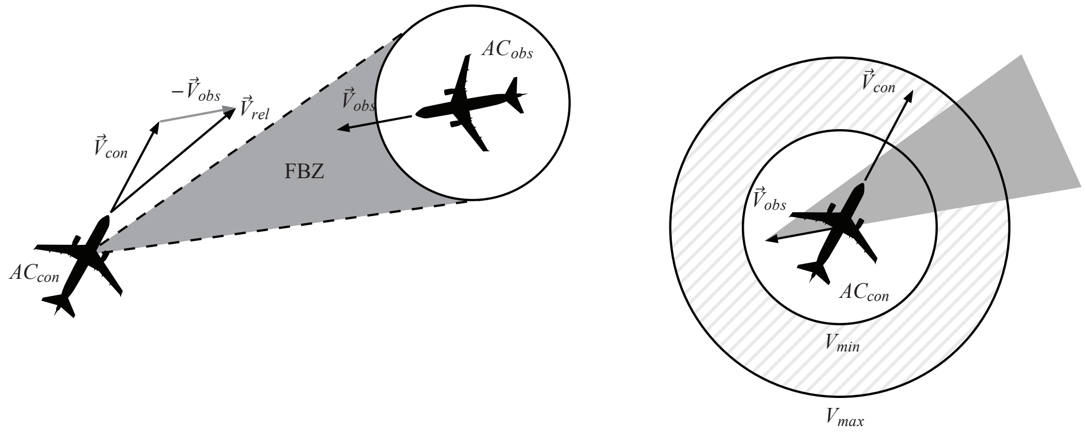
Table 2-5 shows a list of objective measures that can be used to determine the performance of the corresponding metric.

Table 2-5: Measures corresponding to the ATCo competences metrics

	Solving conflicts	
1	Level changes vs vector/speed changes	Number of level, heading and speed changes
2	Minimize the number of aircraft to move	Number of level, heading and speed changes
3	Cut off at nearest available level on climb	Altitude of aircraft; Vertical speed of aircraft; Position of aircraft; Instructed altitude of aircraft
4	Cut off at highest possible level on climb	Altitude of aircraft; Vertical speed of aircraft; Position of aircraft; Instructed altitude of aircraft
5	Descend to nearest available level	Altitude of aircraft; Vertical speed of aircraft; Position of aircraft; Instructed altitude of aircraft
6	Assign the only level available	Altitude of aircraft; Vertical speed of aircraft; Position of aircraft; Instructed altitude of aircraft
7	For crossing conflicts, turn slower aircraft behind	Heading of aircraft; Speed of aircraft; Position of aircraft; Altitude of aircraft
8	For same track conflict, turn faster aircraft direct to route in front of slower aircraft	Heading of aircraft; Speed of aircraft; Position of aircraft; Altitude of aircraft
	R/T	
9	The way of R/T use to keep the workload low	Number of instructions
10	Consistency in the type of instructions	Number of type of instructions
11	Minimize the amount of errors in R/T	Number of errors in the instructions
	Safety	
12	Maintain separation minimums	Number of mid-air collisions, LOS, conflicts; Time in conflict
13	Use sufficient safety buffers	Average distance at CPA; Average Time to CPA; Time between conflict detection and action.
14	Errors in using procedures	Number of errors in the use of procedures
	Efficiency	
15	Minimize additional track miles flown	Total additional track miles flown compared to a direct undisturbed route
16	Minimize delay time	Total delay time compared to a direct undisturbed route
17	Create an expeditious flow of traffic	Outflow of traffic in the sector
18	Moment of traffic handling	Flown track miles at each given instruction

Table 2-6 shows a list of measures corresponding to the ATC structural elements metrics from Table 2-4. The measures from this table will have an influence on the measures from ATCo competences in Table 2-5.

Considering the availability of solution space, this metric can be interpreted in two ways. At first, the Solution Space Diagram (SSD) area can be used as a measure (Figure 2-4). The solution space is the space around the aircraft bounded by its minimum and maximum velocity. This solution space can be graphically represented in the SSD. This 2D SSD covers all possible heading/velocity combinations in which the aircraft can safely move within the sector and all possible heading/velocity combinations in which the aircraft is on a conflict course with another aircraft (Mercado Velasco et al., 2010). Another way to look at the availability of the solution space is to look at the solution space of the whole sector, instead of the solution space of individual aircraft. This is highly linked to the air traffic density in the sector, because with more aircraft inside the sector, less solution space is available. The number of aircraft is a measure for the air traffic density inside the sector.



(a) Plan view of the traffic scenario

(b) Solution Space Diagram for the controlled aircraft. The gray striped SSD area is the available solution space.

Figure 2-4: Solution Space Diagram area of the controlled aircraft (adapted from Mercado Velasco et al.)

As stated in Section 2-3, the DCPA and the TCPA are important values when looking at the CPA. To determine the conflict geometry, the heading of both aircraft in conflict needs to be known. For the determination of the speed difference, the speed of both aircraft is used as a measure. Furthermore, the amount of aircraft that change levels inside the sector has an influence, which is measured by the number of climbing and descending aircraft.

The decisions of the ATCo are determined by the type of aircraft. The type of aircraft gives the ATCo an indication of the performance of an aircraft in terms of, for example, maneuverability and climb/descent rate. Furthermore, the maximum and minimum speed of the aircraft has an influence in the decision making of an ATCo. As a measure for the use of current traffic patterns, the number of traffic patterns, the amount of aircraft per pattern and the variation in traffic pattern can be used.

Finally, when variability in procedures is present in for example the flight level of aircraft that leave the sector, then the altitude of aircraft that leave the sector are used as a measure.

Table 2-6: Measures corresponding to the ATC structural elements

	Traffic volume & density	
1	Availability of solution space	SSD area, Sector solution space area
2	Air traffic density	Number of aircraft
	Position & clearances	
3	Closest point of approach	Time to CPA, Distance to CPA
4	Conflict geometry	Heading of aircraft
5	Speed difference	Speed of aircraft
6	Amount of aircraft changing levels	Number of climbing aircraft, Number of descending aircraft
	Traffic mix & performance	
7	Type of aircraft	Type of aircraft, Maximum/Minimum speed of aircraft
	Traffic flows	
8	Use of current traffic patterns	Number of traffic patterns, Number of aircraft per pattern, Variation in traffic pattern
	Procedures	
9	Variability in procedures	Altitude of aircraft leaving the sector

Chapter 3

Machine Learning

This chapter describes the results from the literature study about Machine Learning (ML) and ML techniques. Section 3-1 describes the definition of ML and gives an overview of the ML types and applications. One of these applications is cluster analysis which is discussed, together with hierarchical clustering, in Section 3-2. Finally, Section 3-3 describes the selection and extraction of features from the dataset.

3-1 Definition

ML is a field within Artificial Intelligence (AI) that gives computer algorithms the ability to learn from experience, learn by example and learn by analogy (Negnevitsky, 2011). There are three main types of learning: unsupervised learning, supervised learning and reinforcement learning (Figure 3-1) (Russell & Norvig, 2010). In unsupervised learning the algorithm is subjected to inputs without letting the algorithm know what the desired output is. With unsupervised learning the algorithm is able to learn certain patterns in the input data with help of an application called clustering. Other applications of unsupervised learning are dimensionality reduction and density estimation. Dimensionality reduction simplifies input data and density estimation is used to find the statistical distribution of input data. In supervised learning the algorithm is subjected to input-output pairs and learns a function that maps input to output. Applications of supervised learning include classification, which classifies observations (from input data) into one of two or more classes, and regression where the outputs are continuous instead of discrete. In reinforcement learning the algorithm learns by getting rewards or punishments from the actions it performs (Russell & Norvig, 2010).

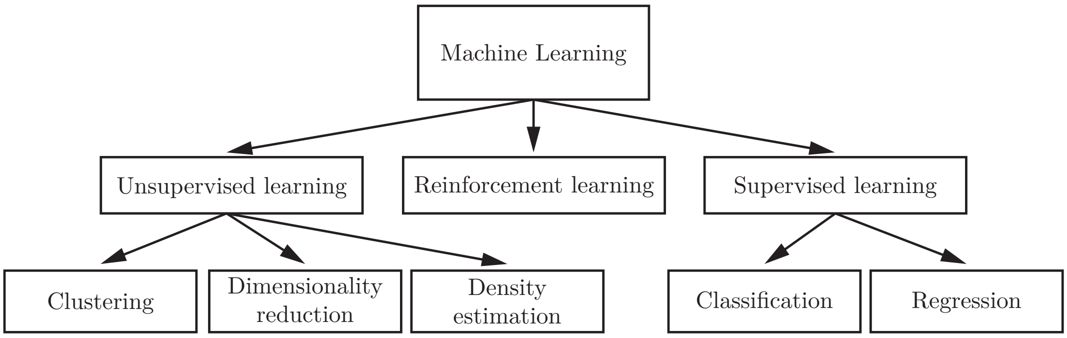


Figure 3-1: Overview of the machine learning types and applications

3-2 Cluster Analysis

Cluster analysis is an exploratory data analysis technique that divides objects into individual groups or clusters. The objects within the same cluster show resemblance with each other and differ in some respects from objects in other clusters (Everitt et al., 2011). As there is no specific output known, the algorithm is mainly used to get insight into the input data.

Within the clustering techniques, clustering can be generally divided into two types: hard-clustering (non-fuzzy clustering) and soft-clustering (fuzzy clustering). With hard-clustering each object belongs to an individual cluster completely or does not belong to a cluster at all. In soft-clustering the likelihood or probability that an object belongs to any cluster is determined. This means that objects can be assigned to multiple clusters.

To deal with a small number of samples in a dataset, it is desired to directly get information about the relationship between all the samples. A good clustering method that achieves this desire is a hard-clustering technique called hierarchical clustering. Since there is a small number of samples in the dataset, hierarchical clustering will be used in this research.

3-2-1 Hierarchical Clustering

In hierarchical clustering the data is organized in a hierarchical structure according to the distance matrix (Xu & Wunsch Ii, 2005). This structure is usually visualized in a dendrogram or binary tree. Looking at a dendrogram, the root node represents the whole dataset, while each leaf node represents a single data object (Figure 3-2). The nodes in between the root and leaf nodes represent how close the objects are to each other. The height of a dendrogram represents the distance between an object and a cluster or a pair of objects and or clusters. When cutting the dendrogram (setting a threshold) at different levels, a visual representation is created for the potential data clustering structures (Xu & Wunsch Ii, 2005). Figure 3-2 shows a threshold of three clusters.

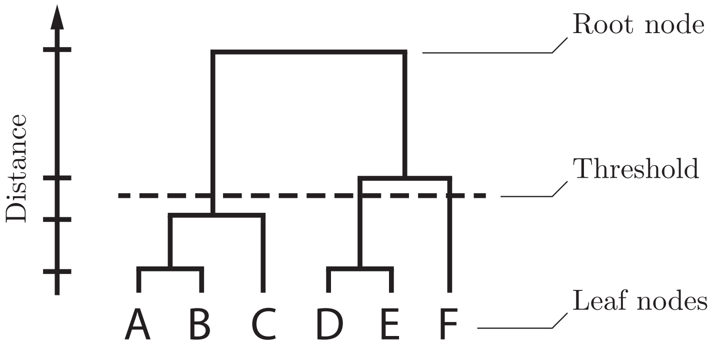


Figure 3-2: An example of a dendrogram including the elements

Within hierarchical clustering there are two methods of clustering: agglomerative and divisive. Agglomerative clustering is a ”bottom up” method which starts with N clusters containing a single data object each (Xu & Wunsch Ii, 2005). In the process that follows the individual clusters are merged which finally leads to one single cluster. Divisive clustering is a ”top down” method that starts as a single cluster containing all the data (Xu & Wunsch Ii, 2005). In the process that follows the clusters are divided until there are only clusters containing a single data object.

The advantages of using hierarchical clustering is that it outputs a hierarchy structure that is more informative compared to the unstructured set of clusters which is returned by a flat clustering algorithm, like K-means clustering (Manning et al., 2008). Furthermore, hierarchical clustering does not need a prespecified number of clusters (Manning et al., 2008). However, the advantages of hierarchical clustering come at a cost of lower efficiency. Looking at agglomerative clustering, the computational complexity is at least $\mathcal{O}(n^2)$ compared to the linear complexity of K-means clustering (Xu & Wunsch Ii, 2005). For divisive clustering the complexity is even worse with a computational complexity of $\mathcal{O}(2^n)$ (Xu & Wunsch Ii, 2005). Therefore, agglomerative clustering will be used in this research.

Agglomerative clustering can be achieved with the following steps. Figure 3-3 shows an example of hierarchical agglomerative clustering after each run through the steps:

1. Initialize N clusters each containing a single data object. Then, calculate the distance matrix for the N clusters.
2. Search the minimal distance

$$D(C_i, C_j) = \min_{\substack{1 \leq m,l \leq N \\ m \neq l}} D(C_m, C_l)$$

   where D is the distance function in the distance matrix. Then, combine cluster C_i and C_j to form a new cluster.
3. Update the distance matrix by computing the distances between the new clusters
4. Repeat steps 2 and 3 until all the objects are in the same cluster.

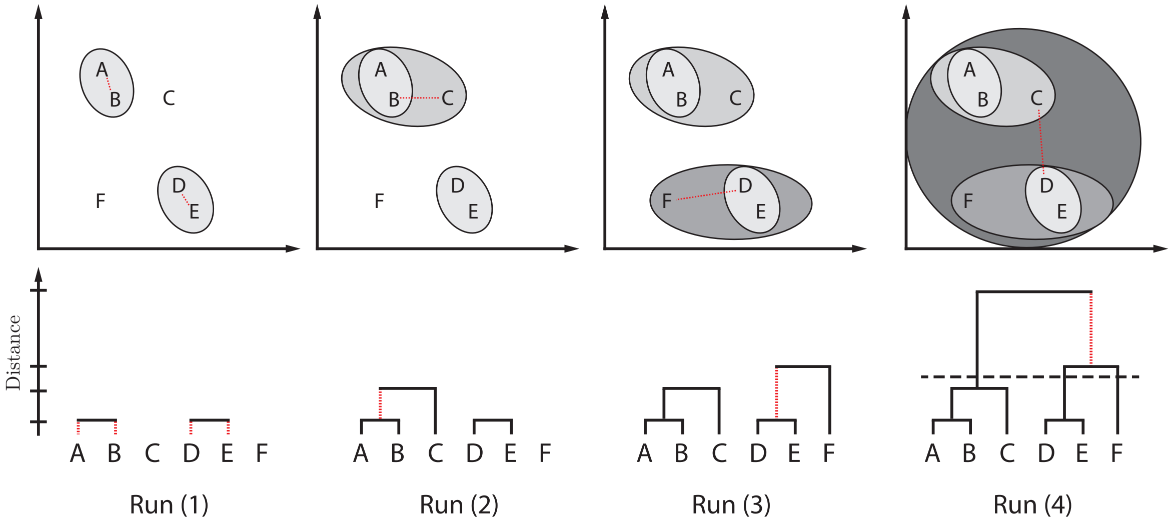


Figure 3-3: Example of hierarchical agglomerative clustering (adapted from Janssen et al.)

The difference between different agglomerative clustering algorithms is determined by the linkage criterion which determines the distance between clusters based on the definition of the distance (Xu & Wunsch Ii, 2005). Examples of commonly used distance metrics are Euclidean distance and Manhattan distance. The used metric has an influence on the shape of the clusters(Xu & Wunsch Ii, 2005). Examples of commonly used linkage criteria are single-linkage, complete-linkage and Ward's method. While single-linkage and complete-linkage use the distance between, respectively, the two closest objects and the two farthest objects in the different clusters, Ward's method tries to keep the total within-cluster sum of squares at a minimum value (Xu & Wunsch Ii, 2005).

By using a certain clustering algorithm, it is desired to get clear and distinct clusters. This means that variance within each cluster should be small and the variance between all clusters should be large. Therefore, the ratio between the between-cluster sum of squares and the total within-cluster sum of squares should be maximized. This ratio is called the F-ratio. Maximizing this ratio can be used as a clustering performance measure (Xu & Wunsch Ii, 2005). No conclusions can be drawn from a single F-ratio value. However, comparing the F-ratio of different clustering results can give information about which clustering result has more distinct clusters. Since Ward's method already tries to keep the total within-cluster sum of squares at a minimum value, using this linkage criterion together with this performance measure can lead to good clustering results.

3-3 Feature Selection and Extraction

Since the data contains many features, the size of the feature set can introduce problems in clustering. These problems include a large computation time, difficulty in interpretation of results and the introduction of the curse of dimensionality (Xu & Wunsch Ii, 2005). The curse of dimensionality describes that when the dimensionality is high enough, the distance between the nearest points is no different from that of other points (Beyer et al., 1999). Distance based clustering algorithms are therefore no longer effective when using high dimensional data.

The idea behind feature selection is that the dataset contains features that are redundant or irrelevant and that selection will remove these types of features. Feature selection is often used in a dataset with many features and a few numbers of observations. Redundant features are features that are already described by other features and therefore provide no additional information. Irrelevant features do not provide any useful information to the clustering method and even negatively impact the clustering results (Kaufman & Rousseeuw, 1990). Irrelevant features will lead to a lot of random terms in the distances and therefore hide the information from the relevant features.

Feature selection techniques can be divided into three different methods: filter methods, wrapper methods and embedded methods (Stańczyk & Jain, 2015). Filter methods gives a ranking to each feature in the subset and the user can select the features based on the ranking. The advantage of these method is that it has a relatively low computation time, but these methods do not consider the relationship between features. Wrapper methods use a model that determine the best performing set of features from all features. The advantage over the filter methods is that wrapper methods do consider the relationship between features, but it can have a relatively high computation time. Furthermore, there is a possible risk of overfitting when the number of observations is not sufficient. Embedded methods select features as part of the model construction process. Embedded methods combine the advantages of both the filter and the wrapper method, but the use of an embedded method highly depends on the used learning model.

Feature extraction is different from feature selection. In feature extraction new features are created from existing features (Xu & Wunsch Ii, 2005). A commonly used feature extraction method is Principal Component Analysis (PCA). PCA reduces the data with use of the principal components of the data. The first principle component describes as much of the variability in the data as possible, and each next component describes as much of the remaining variability as possible (Xu & Wunsch Ii, 2005). With this method the number of dimensions can be reduced. The downside is that the original features cannot be extracted from an PCA. Therefore, the impact of each individual feature or set of features cannot be determined. However, PCA can be used to reduce the dimensionality of the data in such extent that the data can be made visual.

Chapter 4

Preliminary Analysis

With use of the literature survey described in Chapter 2 and Chapter 3 a preliminary analysis is conducted to give practical insight in the theory from the literature. The focus in this preliminary analysis is mainly on the clustering instead of the selection of features. First, a conceptual design is made that describes the process in getting the best feature set which can cluster the ATCos in the different expertise groups (Section 4-1). The data needed for the preliminary analysis is analyzed in Section 4-2, in which the relevant features are extracted from the data. The analyzed data is clustered in Section 4-3 and the results are shown in Section 4-4. Finally, the results from the data clustering are discussed in Section 4-5.

4-1 Conceptual Design

This section describes the conceptual design of the process, in finding a set of features that best describes the different ATCo expertise groups. The feature selection process is shown in Figure 4-1.

From the literature survey, measures are selected which corresponds to metrics that assess the ATCo competences. These competences are influenced by ATC structural elements to which measures are linked to. This results in a set of measures, or features, which are extracted from the dataset. This pre-selection of features from the literature results in a smaller search domain and therefore lower computation time for the feature selection wrapper.

The feature selection wrapper selects from the pre-selected features a subset of features which leads to the most distinct clusters describing the different ATCo expertise groups. First, an initial subset from the pre-selected features is constructed. The wrapper loop uses hierarchical agglomerative clustering to cluster the subset data and evaluates the distinctiveness of the formed clusters. This distinctiveness is used as a fitness criterion for the genetic algorithm. The fitness criterion describes how good the performance is of this feature subset. Based on the fitness evaluation the genetic algorithm creates a new subset of features which is, again, used as an input for the clustering algorithm to determine the performance of this subset. The genetic algorithm creates new subsets based on the current best performing subset and when the new subset is better than the current best performing subset, the new subset becomes the best performing subset. After a stop criterion is reached, the output of the wrapper is the best performing subset of features.

The best performing subset is subjected to a sensitivity analysis. A sensitivity analysis does not only reveal the robustness of the subset, but also can reveal if overfitting has occurred in the feature selection wrapper. Finally, conclusions can be drawn based on the feature selection process.

In this preliminary analysis the focus is mainly on a part of the data analysis, where the data is visualized and features are gathered from the data (Section 4-2), and data clustering, where the gathered features are clustered using hierarchical agglomerative clustering (Section 4-3).

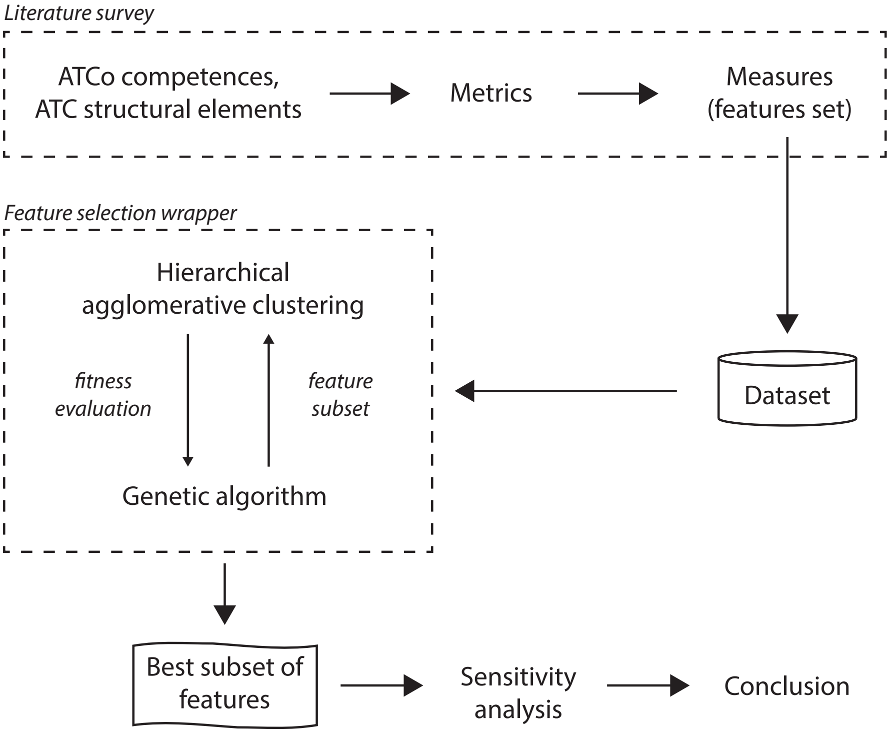


Figure 4-1: Flow diagram of the feature selection process

4-2 Data Analysis

In this section the data from the dataset is analyzed. The dataset consists of data from four different air traffic scenarios solved by ten different ATCos. Four participants were retired ATCos (the professional group) and six participants completed a multiple day extensive ATC-course and/or had worked as a researcher in the ATC field (the ATC course/research group). Each of the participants solved scenarios in a sector that can be compared to the AMS ACC South Sector. At each timestamp, the information of all aircraft, like position, altitude, speed and heading is recorded. Furthermore, the communication between the participant and the aircraft, by means of command inputs, is recorded, and a subjective workload rating is given by each participant.

Looking at the available data, not everything that is recorded could be used to determine the experience of the ATCo or could have an influence on the ATCo. The *use of systems*, *coordination* and *teamwork* are part of the competences from the ACoPOS model shown in Figure 2-2 in Chapter 2, but no data is recorded. Likewise, influences from *systems*, *airport & runways*, *flight plans*, *the team*, *weather conditions* and *emergency situations* are also not in the data. Furthermore, the influence of *airspace & sector* will not be measurable, because there is no change across the scenarios. Although these ATCo competences and ATC structural elements could have an influence in the determination of the expertise level, these competences and elements are not assessed in this research. Therefore, only the selected ATCo competences and ATC structural elements from Table 2-1 from Chapter 2 will be assessed in this research.

Looking at the moment of traffic handling by ATCos it is expected that more experienced ATCos will give a level, heading or speed change far before the aircraft leaves the sector (Table 2-3 from Chapter 2). To visualize this, scatter plots are made of the positions of aircraft when a level, heading or speed command is given to that particular aircraft (Figure 4-2). The scatter plots are separated per expertise group and per scenario. The aircraft in these plots were flying from waypoint AZUL and BLIP, south of the sector, and were merged inside the sector into a single traffic stream towards waypoint MIFA, north of the sector. These traffic streams are chosen in this analysis, because merging two streams can cause potential conflicts which need to be solved by using level, heading and speed commands.

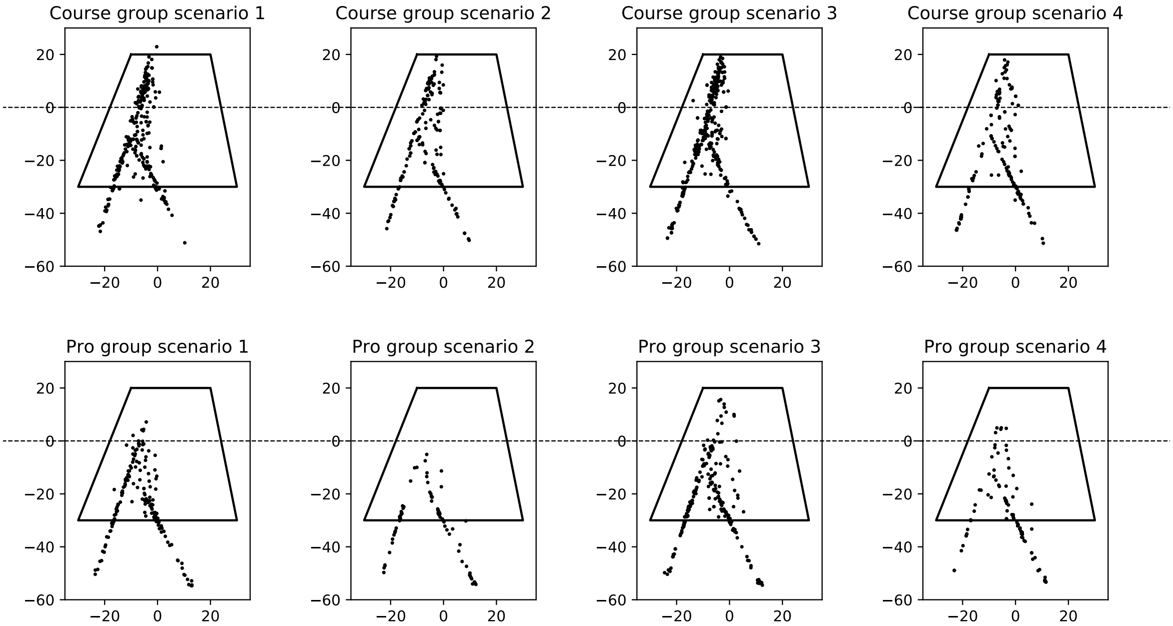


Figure 4-2: Positions of aircraft when a level, heading or speed command is given to that particular aircraft. The scatter plots are separated per expertise group and per scenario.

For the ATC course/research group, it can be seen that a large portion of their commands is given above the dashed line compared to the professional group. Table 4-1 shows the percentage of the given commands that are above the dashed line in Figure 4-2 per expertise group and per scenario. There is a clear difference between the two expertise groups. Therefore, by visually inspecting the data, it is reasonable to think that there is a difference in the moment of traffic handling between the expertise groups.

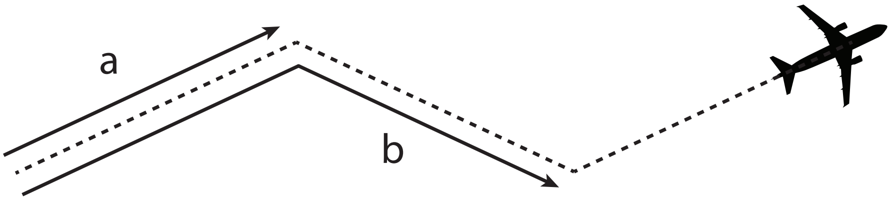


Figure 4-3: Flown track of an aircraft that has been subjected to two heading changes. Length *a* represents the flown track miles before the first heading change. Length *b* represents the flown track miles before the second heading change.

$$\sum\nolimits_{\text{squared track miles}} = a^2 + b^2 + ... \tag{4-1}$$

To measure to what extent this metric is expressed, for each aircraft in the scenario the sum of the squared track miles when a level, heading or speed command is given is used (Figure 4-3 and Equation 4-1). For each ATCo, each scenario, each aircraft and each command type the sum of the squared track miles, when a particular command type is given, is obtained. These sums are taken together to get a single sum of squared track miles for each ATCo, each scenario and each command type. It is expected that the professional group has a lower overall sum of squared track miles compared to the ATC course/research group. These measures are used for the data clustering in Section 4-3.

Table 4-1: Percentage of the given commands that are above the dashed line from Figure 4-2 per expertise group and per scenario

	Scenario			
	1	2	3	4
Course group	33%	31%	32%	25%
Professional group	3.0%	0.0%	6.2%	5.8%

Looking at the type of instructions given by ATCos it is expected that more experienced ATCos are more consistent in the type of given instructions (Table 2-2 from Chapter 2). To measure this, the number of Direct (DCT), Executive Flight Level (EFL), Heading (HDG) and Speed (SPD) commands are gathered from the data. These measures are used for the data clustering in Section 4-3.

Table 4-2 shows the metrics with corresponding measures discussed in this section that are used in the data clustering in Section 4-3.

Table 4-2: Metrics and measures for the preliminary analysis

1	Moment of traffic handling	Sum of squared track miles for level commands Sum of squared track miles for heading commands Sum of squared track miles for speed commands
2	Consistency in type of instructions	Number of DCT commands Number of EFL commands Number of HDG commands Number of SPD commands

4-3 Data Clustering

In this section the data from the data analysis is clustered based on the measures, or features, from Table 4-2. Hierarchical agglomerative clustering is used, because this clustering method outputs a hierarchy structure, a dendrogram, that is more informative compared to unstructured set of clusters, which is returned by a flat clustering algorithm, like K-means clustering. Furthermore, agglomerative clustering is better in terms of computational complexity, compared to divisive clustering as stated in Chapter 3.

As stated in Chapter 3, Ward's method will be used as the linkage criterion and the ratio between the between-cluster sum of squares and the total within-cluster sum of squares will be used as a performance measure. Since Ward's method already tries to keep the total within-cluster sum of squares at a minimum value, it is beneficial for the maximization of the performance measure.

From the measures from each metric in Table 4-2, the ratio of each measure to the total sum of all measures of that metric is calculated. After the ratios have been calculated, the entire set of measures is standardized to get all the measures on the same scale with zero mean and unit variance. Section 4-4 shows the results from the data clustering.

4-4 Results

In this section, the results are shown of the data clustering. Figure 4-4 shows the clustering results of the metric 1: *"Moment of traffic handling"*. Figure 4-5 shows the clustering results of the metric 2: *"Consistency in type of instructions"*. Figure 4-6 shows the clustering results of using both metrics. The ATCos from the professional group are shown in black and the ATCos of the ATC course/research group are shown in gray. Furthermore, since it is desired to have two distinct clusters, the dendrograms are cut, such that two clusters emerge which are boxed with dashed lines.

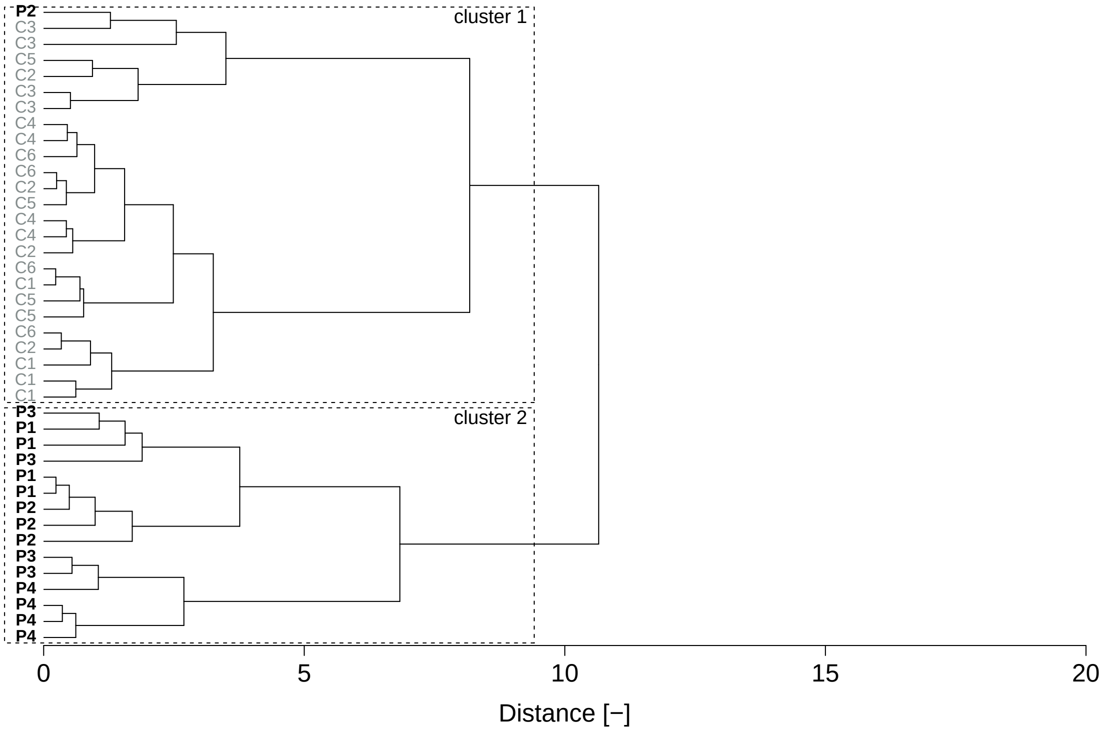


Figure 4-4: Dendrogram of the metric 1: *"Moment of traffic handling"* with two clusters showing the ATC course/research ATCos in gray and the professional ATCos in black

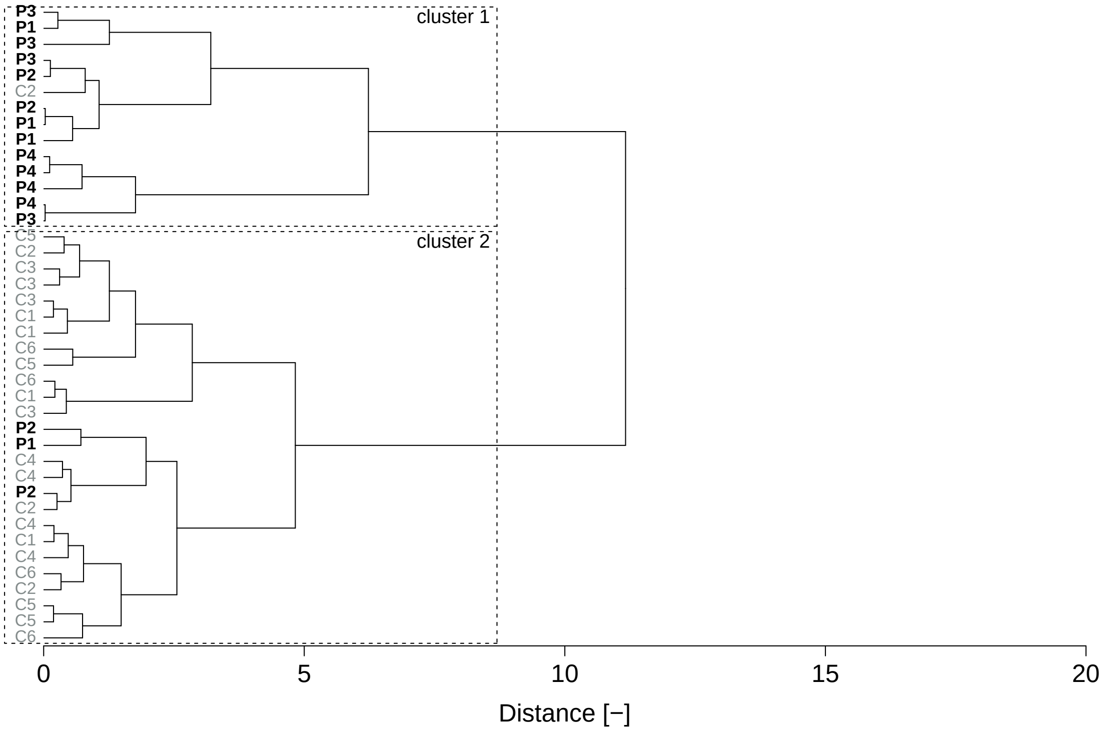


Figure 4-5: Dendrogram of the metric 2: *"Consistency in type of instructions"* with two clusters showing the ATC course/research ATCos in gray and the professional ATCos in black

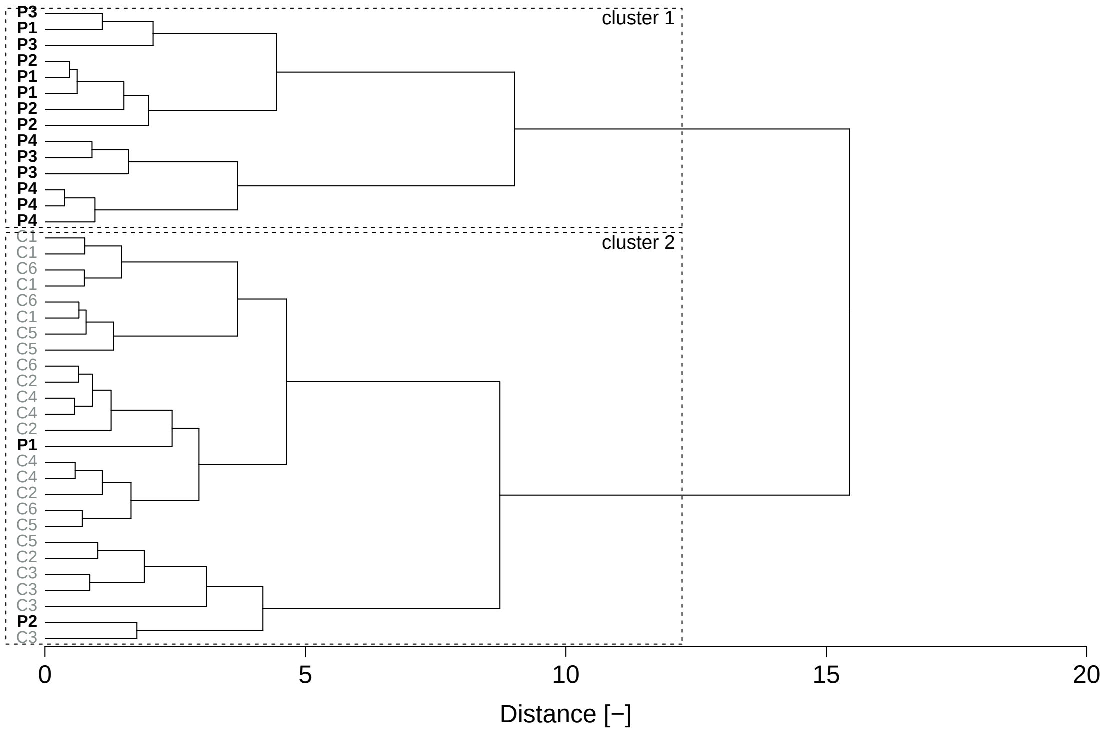


Figure 4-6: Dendrogram of the both metrics (*"Moment of traffic handling"* and *"Consistency in type of instructions"*) with two clusters showing the ATC course/research ATCos in gray and the professional ATCos in black

Table 4-3 shows the performance parameters of the clustering results. These parameters include the Total Sum of Squares (TSS), the Within-cluster Sum of Squares (WSS), the Between-cluster Sum of Squares (BSS) and the F-ratio, which is the ratio between BSS and WSS.

Table 4-3: Performance parameters of the clustering results of both individual metrics and combined metrics

	Metric		
	1	2	1+2
Total sum of squares (TSS)	156	117	273
Within-cluster sum of squares (WSS)	99.2	54.7	154
Between-cluster sum of squares (BSS)	56.7	62.3	119
F-ratio (BSS/WSS)	0.57	1.14	0.78

4-5 Discussion

From the results from Section 4-4 interesting insights can be gathered for future analysis. When looking at the emerging clusters of all three dendrograms, it can be seen that the most ATCos tend to cluster with ATCos from the same expertise group. Therefore, showing that it is indeed possible to cluster the data into the two different expertise groups with use of these metrics. The F-ratios of all dendrograms show which metrics result in the most distinct clusters (Table 4-3). This shows that the data from the metric 2 (Figure 4-5) result in the most distinct clusters.

Although Figure 4-5 shows the best performing metric, it also shows that one cluster does not only contain professionals and one cluster does not only contain ATC course/research ATCos. Furthermore, the TSS, WSS and BSS of metric 1+2 is the result of a summation of the TSS, WSS and BSS from metric 1 and metric 2, but no such relationship can be established about the composition of the clusters. Therefore, during feature selection it is possible that badly composed clusters can have a relative large F-ratio. This raises the question that for future analysis possibly an additional performance parameter is needed to create clusters, containing only ATCos of a single expertise group.

Looking at the F-ratio of metric 1+2 (Table 4-3) it shows that it is performing better than just metric 1, but worse compared to just metric 2. This shows that using more metrics to cluster the data does not always lead into better performance. Furthermore, it must be noted that although there are 7 measures (3 measures from metric 1 and 4 measures from metric 2) used for the clustering of metric 1+2, the measures of each metric depend on each other, because the ratio is taken. Removing a measure can therefore lead into completely different ratios. This must be considered with the feature selection in future analysis.

Chapter 5

Conclusion

In this preliminary thesis a first analysis is given about how a clustering algorithm can be used to determine a set of objective measures, based on good control behavior, that accurately describe an air traffic controller's expertise level. At first, a literature survey was conducted on the ATCo competences and assessment. In this chapter, relevant competences are discussed, and corresponding metrics and measures are linked to each competence. Furthermore, relevant ATC structural elements, that have an influence on the decision making of an ATCo, are discussed and corresponding metrics and measures are linked to each element. Secondly, the concept of ML and an application of clustering, hierarchical clustering, is discussed. Furthermore, methods are presented for the selection of features. Finally, a preliminary analysis is conducted to show the capabilities of hierarchical clustering with a part of the dataset.

In this chapter, Section 5-1 describes the preliminary conclusion using the results from the literature study and the preliminary analysis. Section 5-2 describes the steps that need to be taken for future analysis.

5-1 Preliminary Conclusion

This section describes the preliminary conclusions of this preliminary thesis. Answers will be given to the first two subquestion. Furthermore, the third subquestion will be partially answered. To answer the remaining subquestions and the research question future research needs to be conducted.

"Which objective measures can accurately describe good control behavior?"

Table 5-1 shows the objective measures that can accurately describe good control behavior. The objective measures are categorized into measures from ATCo competences and measures from ATC structural elements. Some measures are in both categories, because they both serve as an influence in the decision making of an ATCo and to determine to what extent a metric is used by an ATCo. There is a possibility that more measures exist, but this depends on the metrics that are formed when visually inspecting the data in further analysis.

Table 5-1: Objective measures that can accurately describe good control behavior. Categorized into measures from ATCo competences and measures from ATC structural elements.

ATCo competences

- Altitude of aircraft
- Average distance at CPA
- Average Time to CPA
- Flown track miles at each given instruction
- Heading of aircraft
- Instructed altitude of aircraft
- Number of errors in the instructions
- Number of errors in the use of procedures
- Number of instructions
- Number of level, heading and speed changes
- Number of mid-air collisions, LOS, conflicts
- Number of type of instructions
- Outflow of traffic in the sector
- Position of aircraft
- Speed of aircraft
- Time between conflict detection and action.
- Time in conflict
- Total additional track miles flown compared to a direct undisturbed route
- Total delay time compared to a direct undisturbed route
- Vertical speed of aircraft

ATC structural elements

- Altitude of aircraft leaving the sector
- Distance to CPA
- Heading of aircraft
- Maximum/Minimum speed of aircraft
- Number of aircraft
- Number of aircraft per pattern,
- Number of climbing aircraft and descending aircraft
- Number of traffic patterns
- Solution space area
- Speed of aircraft
- SSD area
- Time to CPA
- Type of aircraft
- Variation in traffic pattern

"Which clustering algorithms can be used to determine a set of objective measures?"

Based on a dataset with a small number of samples and to keep the computational complexity relatively low, hierarchical agglomerative clustering is used as the basis for clustering. The difference between the different hierarchal agglomerative clustering algorithms is determined by the linkage criterion which determines the distance between clusters based on the definition of the distance (Xu & Wunsch Ii, 2005). Table 5-2 shows which linkage criteria and which distance measures can be used to form different clustering algorithms.

Table 5-2: Clustering algorithms that can be used to determine a set of objective measures. Both different linkage criteria and distance measures are shown.

Linkage criteria	**Distance measures**
• Single-linkage	• Euclidean
• Complete-linkage	• Manhattan
• Ward's method	

From the results from the preliminary analysis it can be seen that a hierarchical agglomerative clustering algorithm using Ward's method and a Euclidean distance measure can cluster the data into the professional group and the ATC course/research group.

To measure the performance of the clustering algorithms it is desired to keep the variance within each cluster small and the variance between all clusters large. Therefore, the ratio between the between-cluster sum of squares and the total within-cluster sum of squares (F-ratio) should be maximized.

5-2 Future Steps

Future analysis is needed to answer the remaining subquestions and finally the research question. The preliminary analysis only conducted part of the conceptual design. The next step will be to implement the feature selection wrapper that will select the best subset of features.

Furthermore, since the primary analysis only conducted a part of the metrics and measures, the next step is to conduct all metrics from the literature survey. The different clustering algorithms (linkage criteria and distance measures) need to be tested to get the best performing algorithm conducting all the metrics. A sensitivity analysis needs to be performed to reveal the robustness of the subset and the algorithm and if overfitting has occurred.

An experiment with novices needs to be performed to increase the dataset with new data from a new expertise group (novice group). These novice participants will perform the same experiment (with the same scenarios) as the professional group and the ATC course/research group did.

With the best performing clustering algorithm, best subset of features and the new dataset, the performance is measured of the clustering algorithm when it is subjected to the new unseen data. It is then desired that the chosen clustering algorithm and subset of features will result in three distinct clusters.

Part III

Appendices

Appendix A

Data and Measures

A-1 Conflict Detection

A large part of the preliminary research has been spent on the search for conflict resolution metrics. Table A-1 shows the conflict resolution metrics that were found from the literature. Since the scenarios that the 10 ATCos solved contained a traffic merging problem, it was expected that sufficient conflicts were present in the data. Conflicts were indeed present in each solved scenario, but the amount of conflicts were too small to detect a difference between the expertise groups. Furthermore, the conflict resolutions used were also categorized in the metrics from Table A-1. This means that the little conflicts that were detected are spread even further, so that differences between the expertise groups were even harder to find. Therefore, the *solving conflicts* competences have not been researched further. This competence could be included in the future when the dataset is much larger.

Table A-1: Conflict resolution metrics to assess ATCo competences

	Solving conflicts	
1	Level changes vs vector/speed changes	(Fothergill & Neal, 2013)
2	Minimize the number of aircraft to move	(Fothergill & Neal, 2013)
3	Cut off at nearest available level on climb	(Fothergill & Neal, 2013)
4	Cut off at highest possible level on climb	(Fothergill & Neal, 2013)
5	Descend to nearest available level	(Fothergill & Neal, 2013)
6	Assign the only level available	(Fothergill & Neal, 2013)
7	For crossing conflicts, turn slower aircraft behind	(Kirwan & Flynn, 2001)
8	For same track conflict, turn faster aircraft direct to route in front of slower aircraft	(Kirwan & Flynn, 2001)

A-2 List of Measures

This section shows a list of 59 measures in which a set of measures can be extracted by the genetic algorithm. The name of the measure and a short description is given.

Table A-2: List of 59 measures in which a set of measures can be extracted

	Measure name	**Short description**
1	numOfcmds	*Number of commands*
2	ratio_DCT	*Ratio DCT commands*
3	ratio_EFL	*Ratio EFL commands*
4	ratio_HDG	*Ratio HDG commands*
5	ratio_SPD	*Ratio SPD commands*
6	totalPenalty	*Total trackpenalty*
7	ratio_levelPenalty	*Ratio trackpenalty when using level commands*
8	ratio_headingPenalty	*Ratio trackpenalty when using heading commands*
9	ratio_speedPenalty	*Ratio trackpenalty when using speed commands*
		Mean occupied SSD area at each command
10	pctChangeMeanOccupiedSSDAreaAtCommand	: *Percentage change*
11	signRatioMeanOccupiedSSDAreaAtCommand	: *Ratio of the sign of change*
12	meanMeanOccupiedSSDAreaAtCommand	: *Mean*
13	sdMeanOccupiedSSDAreaAtCommand	: *SD*
14	meanChangeMeanOccupiedSSDAreaAtCommand	: *Mean of change*
15	sdChangeMeanOccupiedSSDAreaAtCommand	: *SD of change*
		Occupied SSD areas
16	meanOccupiedSSDAreas	: *Mean*
17	sdOccupiedSSDAreas	: *SD*
18	maxOccupiedSSDAreas	: *Maximum*
19	minOccupiedSSDAreas	: *Minimum*
		Aircraft time in sector
20	meanAircraftTimeInSector	: *Mean*
21	sdAircraftTimeInSector	: *SD*
		Finished aircraft per logpoint
22	meanFinishedAircraftPerLogpoint	: *Mean*
23	sdFinishedAircraftPerLogpoint	: *SD*

	Measure name	**Short description**
		Outflow (per total logpoints) of aircraft per logpoint
24	maxOutflowPerLogpoint	: *Maximum*
25	maxChangeOutflowPerLogpoint	: *Maximum change*
26	mseChangeOutflowPerLogpoint	: *MSE of change*
27	meanOutflowPerLogpoint	: *Mean*
28	sdOutflowPerLogpoint	: *SD*
		Outflow (per current logpoint) of aircraft per logpoint
29	meanOutflow2PerLogpoint	: *Mean*
30	sdOutflow2PerLogpoint	: *SD*
		Relative distance beween aircraft per logpoint
31	meanSdRelDistancePerLogpoint	: *Mean of SD per logpoint*
32	sdSdRelDistancePerLogpoint	: *SD of SD per logpoint*
33	meanMinRelDistancePerLogpoint	: *Mean of minimum per logpoint*
34	sdMinRelDistancePerLogpoint	: *SD of minimum per logpoint*
35	meanMaxRelDistancePerLogpoint	: *Mean of maximum per logpoint*
36	sdMaxRelDistancePerLogpoint	: *SD of maximum per logpoint*
		Flight level of aircraft to MIFA
37	meanFlightLevelAircraftToMIFA	: *Mean*
38	sdFlightLevelAircraftToMIFA	: *SD*
39	maxFlightLevelAircraftToMIFA	: *Maximum*
40	minFlightLevelAircraftToMIFA	: *Minimum*
		Level changes per aircraft
41	meanLevelChangesPerAircraft	: *Mean*
42	sdLevelChangesPerAircraft	: *SD*
43	maxLevelChangesPerAircraft	: *Maximum*

	Measure name	Short description
		Average TCPA per logpoint
44	meanAverageTcpaPerLogpoint	: *Mean*
45	sdAverageTcpaPerLogpoint	: *SD*
46	maxAverageTcpaPerLogpoint	: *Maximum*
47	minAverageTcpaPerLogpoint	: *Minimum*
		Average TLOS per logpoint
48	meanAverageTlosPerLogpoint	: *Mean*
49	sdAverageTlosPerLogpoint	: *SD*
50	maxAverageTlosPerLogpoint	: *Maximum*
51	minAverageTlosPerLogpoint	: *Minimum*
		Average DCPA per logpoint
52	meanAverageDcpaPerLogpoint	: *Mean*
53	sdAverageDcpaPerLogpoint	: *SD*
54	maxAverageDcpaPerLogpoint	: *Maximum*
55	minAverageDcpaPerLogpoint	: *Minimum*
		Average relative distance between aircraft per logpoint
56	meanAverageRelDistancePerLogpoint	: *Mean*
57	sdAverageRelDistancePerLogpoint	: *SD*
58	maxAverageRelDistancePerLogpoint	: *Maximum*
59	minAverageRelDistancePerLogpoint	: *Minimum*

A-3 Boxplots of Measures

In this section the boxplots of the raw results of the 55 measures are shown in Figure A-1, Figure A-2, Figure A-3 and Figure A-4. This is after the correlated measures were removed. These removed measures are the mean and standard deviation of the finished aircraft per logpoint, the standard deviation of the outflow (per total logpoints) of aircraft per logpoint, and the mean flight level of aircraft flying to waypoint MIFA.

Figure A-1: Boxplots of measure 1 to 16

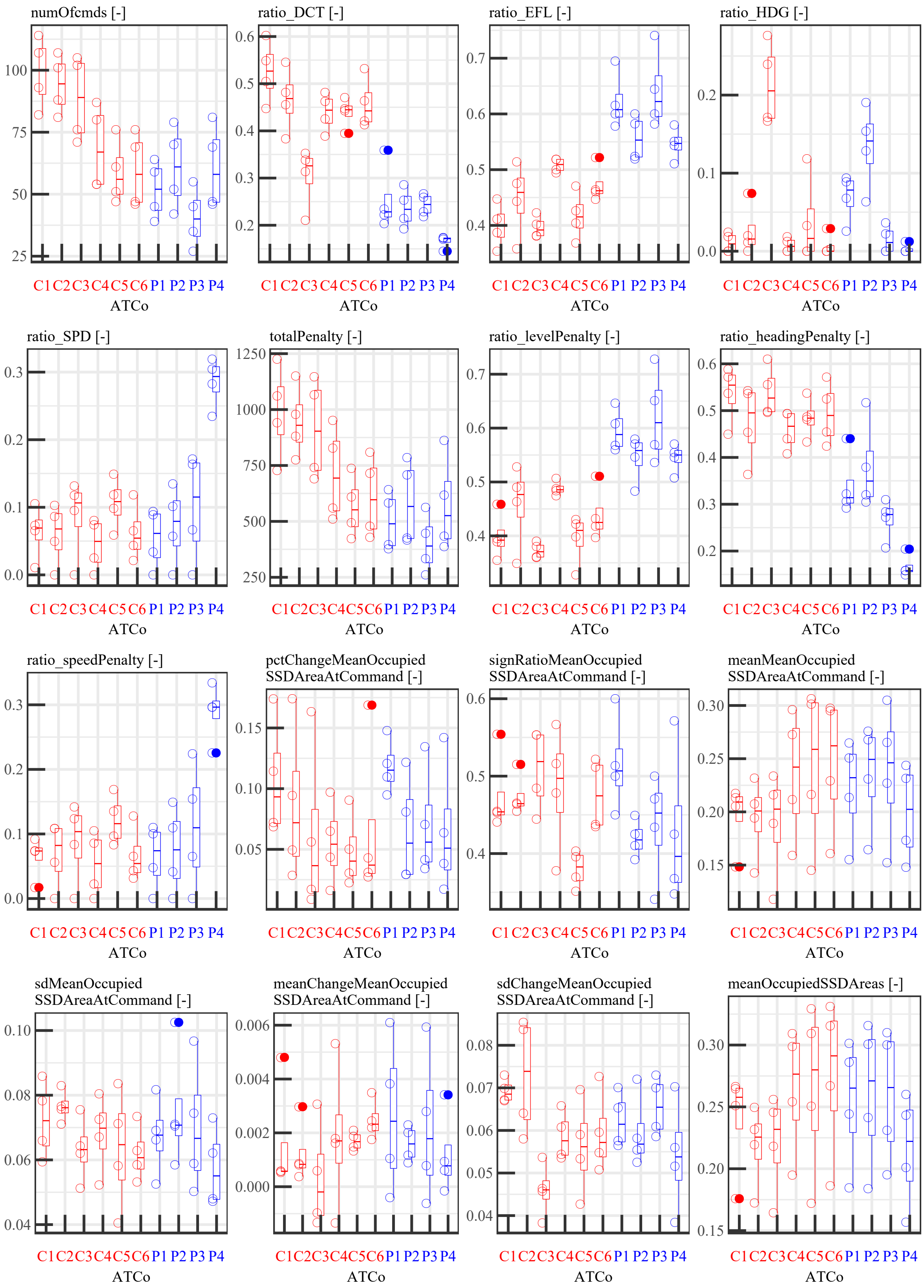


Figure A-2: Boxplots of measure 17 to 32

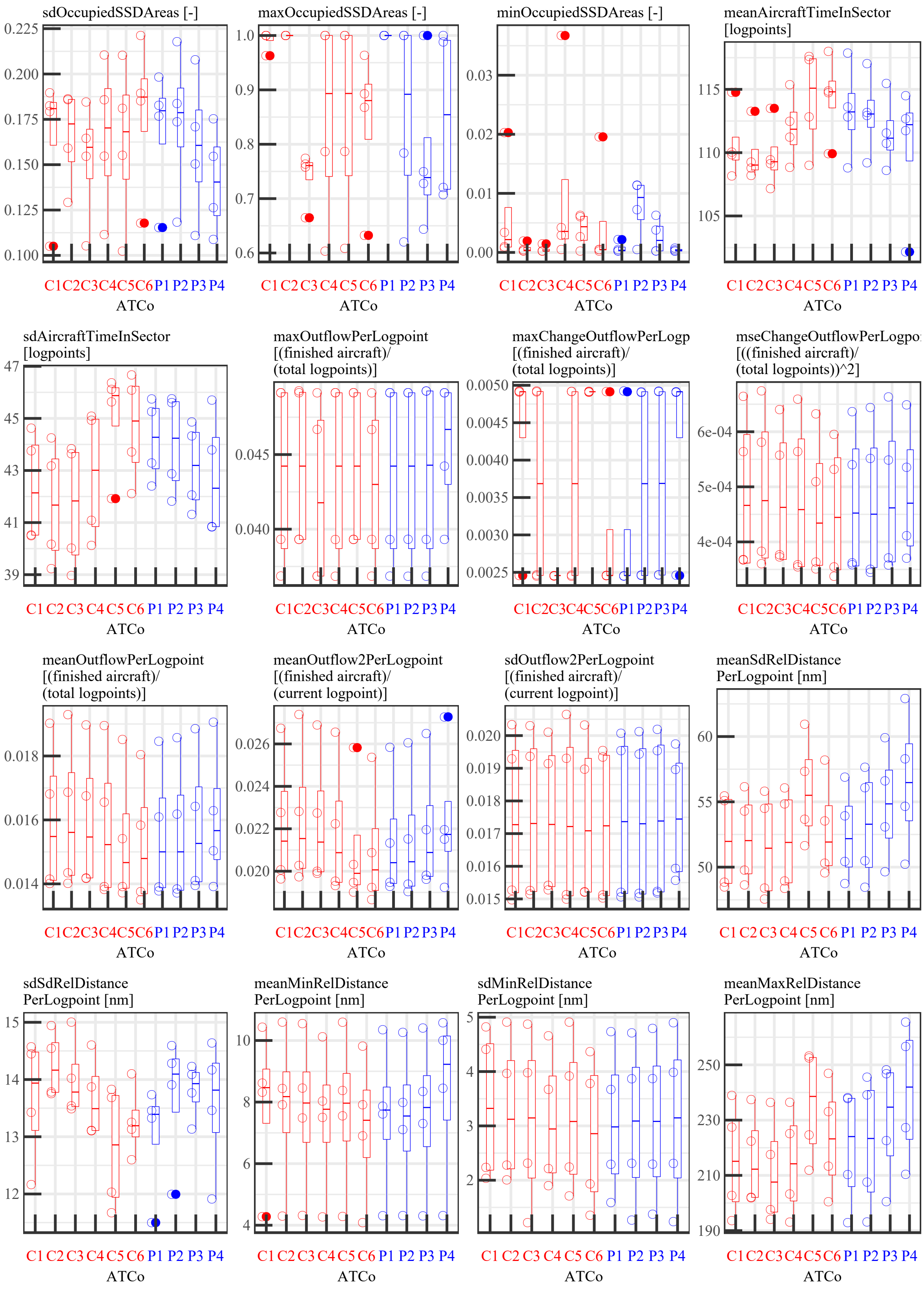


Figure A-3: Boxplots of measure 33 to 48

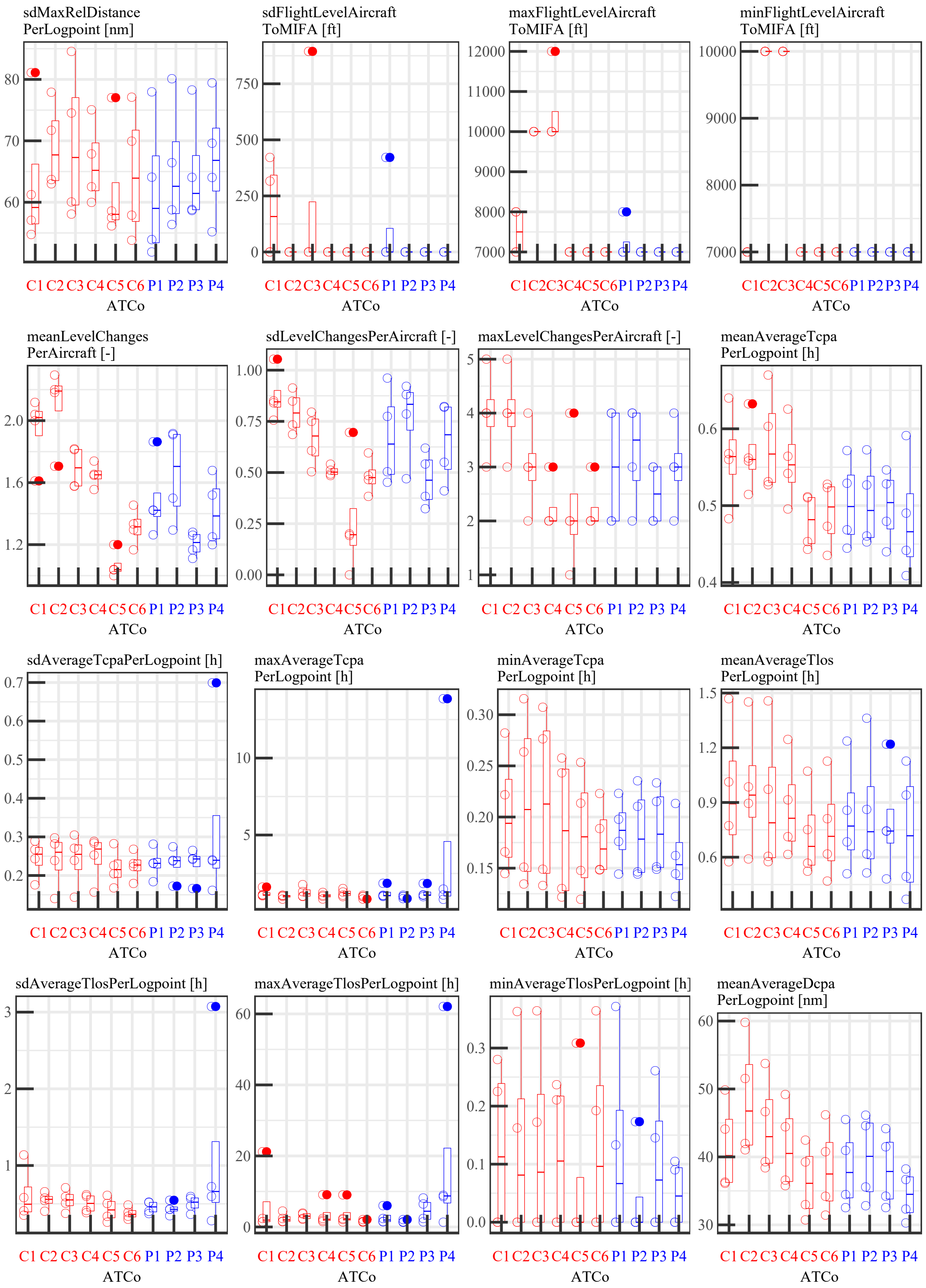
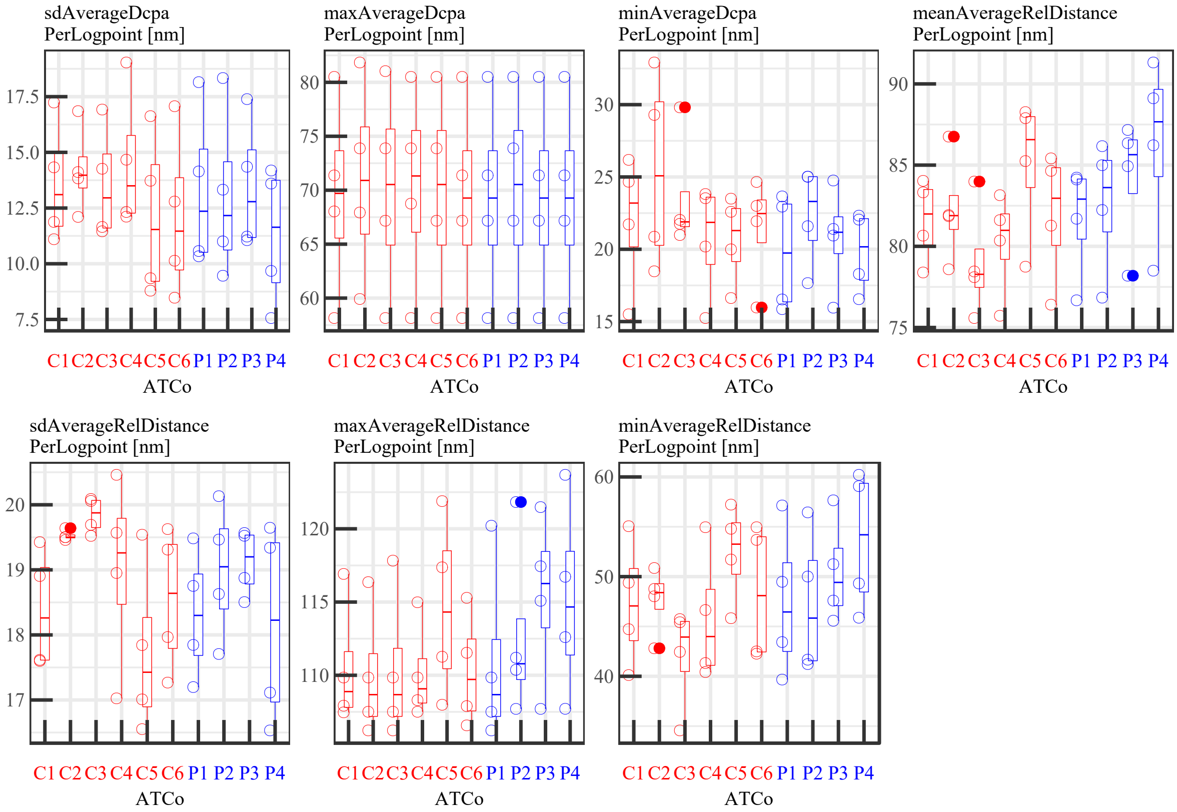


Figure A-4: Boxplot of measure 48 to 55



Appendix B

Genetic Algorithm and Hierarchical Clustering

B-1 Genetic Algorithm Theory

This section describes the theory of hierarchical clustering and a genetic algorithm.

A genetic algorithm is a technique within the concept of the evolutionary computation. Evolutionary computation is based on computational models of natural selection and genetics, and is the evolutionary approach to machine learning (Negnevitsky, 2011). All evolutionary computation techniques simulate evolution by using the biological inspired concepts of selection, mutation and reproduction to explore the search domain (Negnevitsky, 2011).

Genetic algorithms are used to find solutions to optimization and search problems. The advantage of using a genetic algorithm is that it can search in a huge search domain relatively quick. A genetic algorithm does not get stuck in a local optimum, because it uses mutation in the chromosome which is equivalent to a random search in the search domain (Negnevitsky, 2011).

Genetic algorithms are defined as *"a class of stochastic search algorithms based on biological evolution"* (Negnevitsky, 2011, p.222). The genetic algorithm measures the performance of the individual chromosomes based on a fitness function to carry out reproduction. When reproduction takes place, crossover between parts of chromosomes and mutation of individual values within a chromosome take place. After a number of successive reproductions, the result is that lower performing chromosomes will disappear and higher performing chromosomes will excel (Negnevitsky, 2011).

At the start of the algorithm the candidate solutions must be represented by a binary string of a fixed length. This binary string will be the chromosome within the population. An initial chromosome population of size N is randomly generated where each chromosome has the same length as the amount of candidate solutions: $x_1, x_2, ..., x_N$. Furthermore, the crossover probability p_c and the mutation probability p_m is set. Finally, a fitness function is defined to measure the performance of an individual chromosome. The fitness of a chromosome forms the basis for the selection of chromosomes that will be used for reproduction (Negnevitsky, 2011).

After the initialization of the algorithm an iterative process follows according to the following steps (Negnevitsky, 2011). Figure B-1 shows an example of the genetic algorithm after each run through the steps:

1. Calculate the fitness of each individual chromosome in the population:
   $f(x_1), f(x_2), ..., f(x_N)$
2. Select a pair of chromosomes from the current population for reproduction. The pair of chromosomes are selected with a probability based on their fitness. Chromosomes with a high fitness have a higher probability to be selected.
3. Create a new pair of offspring chromosomes from the selected chromosomes in step 2 by using crossover and mutation.
4. Place the created offspring chromosomes in the new population.
5. Repeat step 2 until the size of the new chromosome population has the same size as the initial population.
6. Replace the initial population with the new chromosome population.
7. A new generation was born! Go to step 1 and repeat the process to create the next generation until the termination criterion is satisfied.

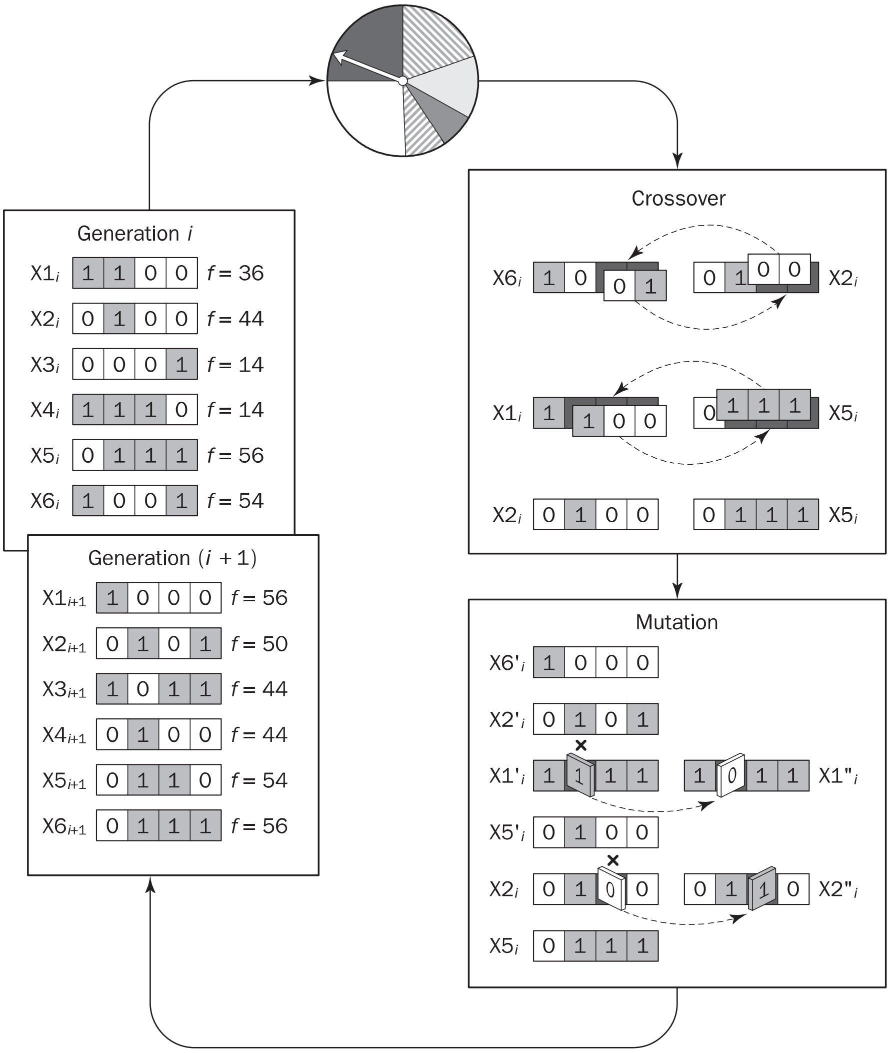


Figure B-1: Example of the GA cycle (adapted from Negnevitsky (2011))

In Figure B-1 a roulette wheel is present, which is a commonly used chromosome selection technique (Negnevitsky, 2011). Each chromosome has its own slice on the roulette wheel and the size of the slice depends on how fit the chromosome is. Therefore, chromosomes that have a higher fitness will have a higher probability to be chosen to create offspring.

B-2 Fitness Functions

Several fitness functions have been tried using *max(BSS/WSS)*. The use of *max(BSS/WSS)* is discussed in the preliminary report (Part II). This subsection shows the advantages and disadvantages of the tried fitness functions.

Fitness functions with *max(BSS/WSS)* (Figure B-2):

1. Does not consider cluster composition.
2. High tendency to pick "sawtooth measures".

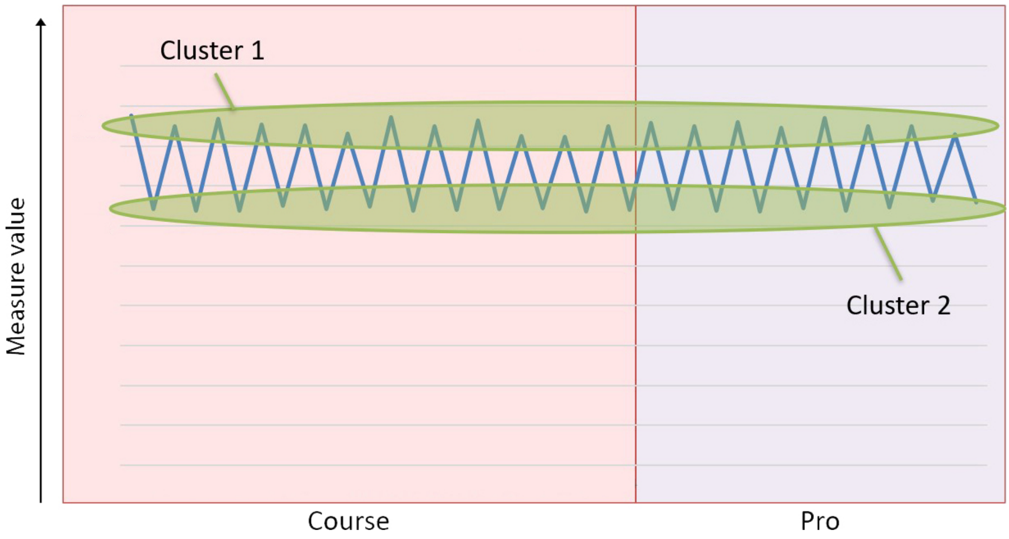


Figure B-2: Cluster phenomenon when the fitness function contains *max(BSS/WSS)*.

Split population into the expertise groups beforehand and calculate *max(BSS/WSS)* (Figure B-3):

1. Does consider cluster composition!
2. No actual clustering is performed, because the desired expertise clusters are already determined beforehand by splitting the population.
3. Picks good performing measures.

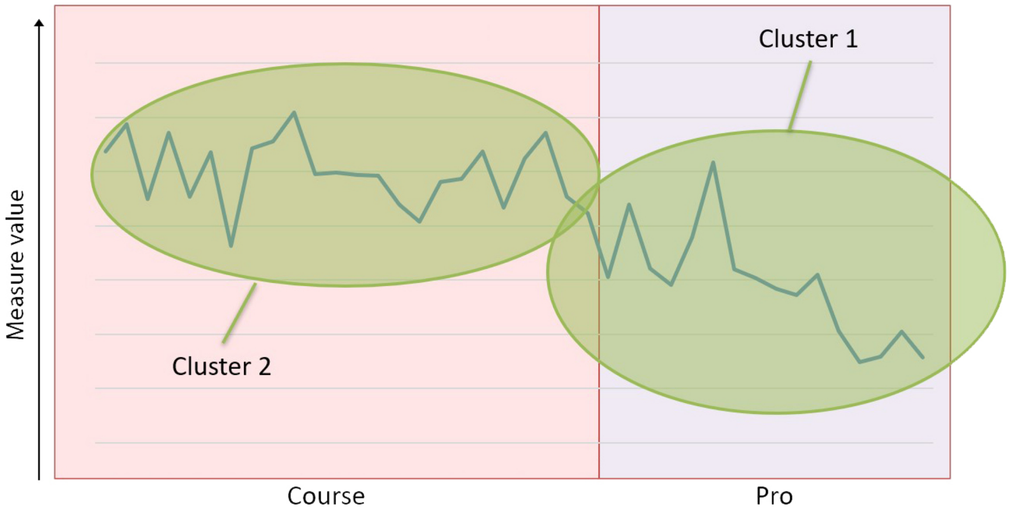


Figure B-3: Cluster phenomenon when *max(BSS/WSS)* is calculated for the individual expertise groups

B-3 PCA Representation

This section shows a principal component analysis (PCA) representation of the clustered data set using the 8 measures from the best set of measures. This representation could be used to get graphical insight in the clustering results. Figure B-4 shows a 3D representation while Figure B-5, Figure B-6 and Figure B-7 show the 2D representation.

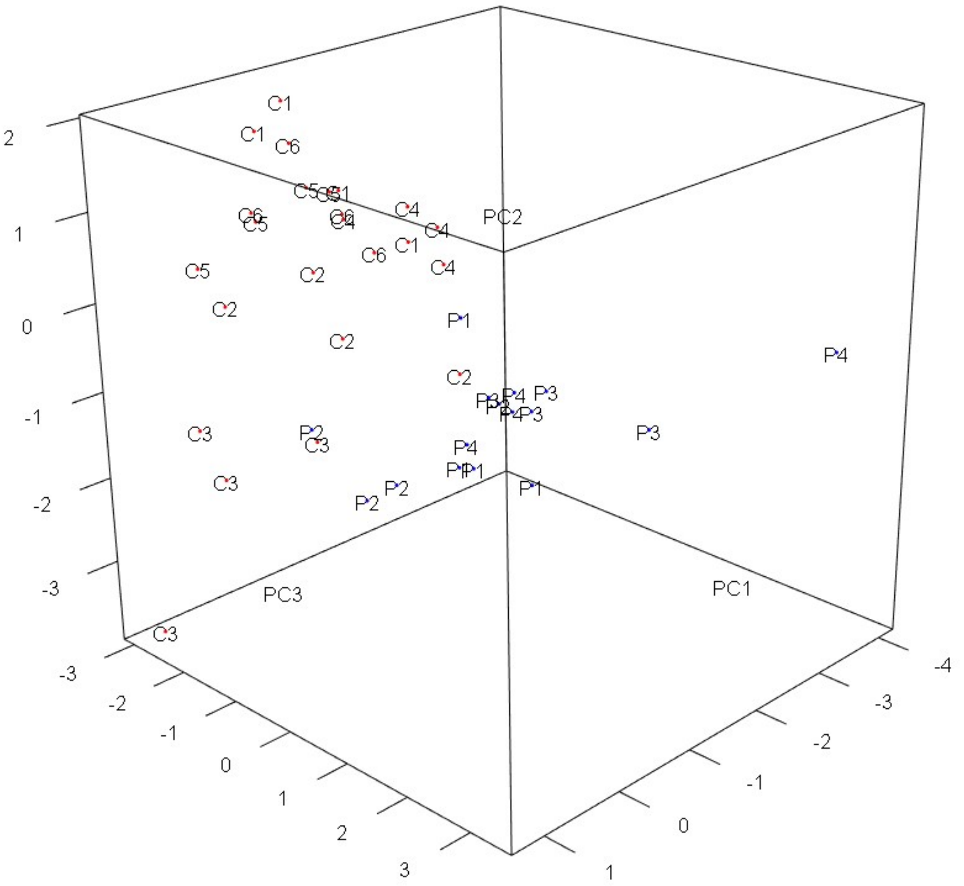


Figure B-4: PCA in 3D of the standardized data [-]

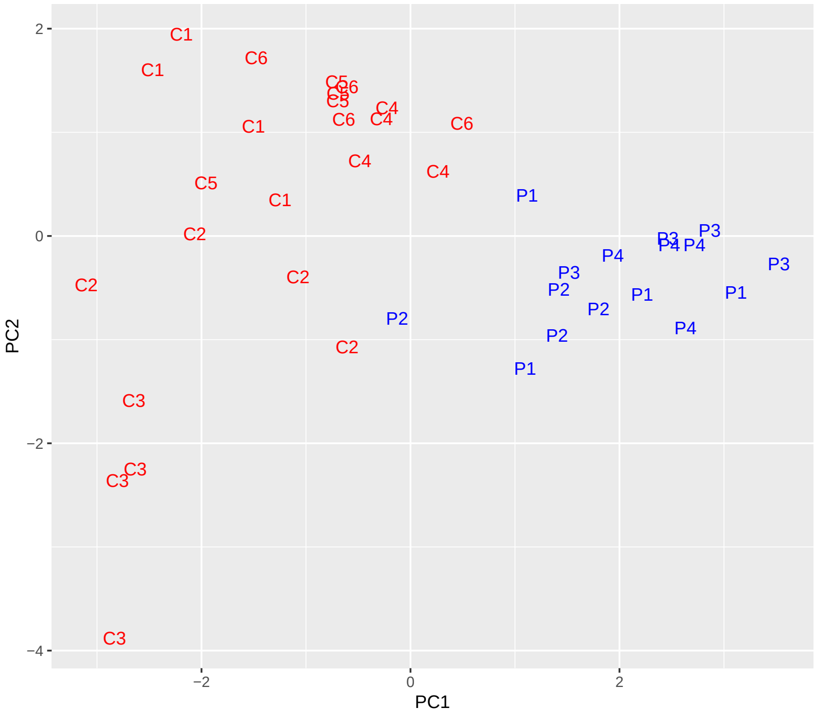


Figure B-5: PCA 1 and PCA 2 of the standardized data [-]

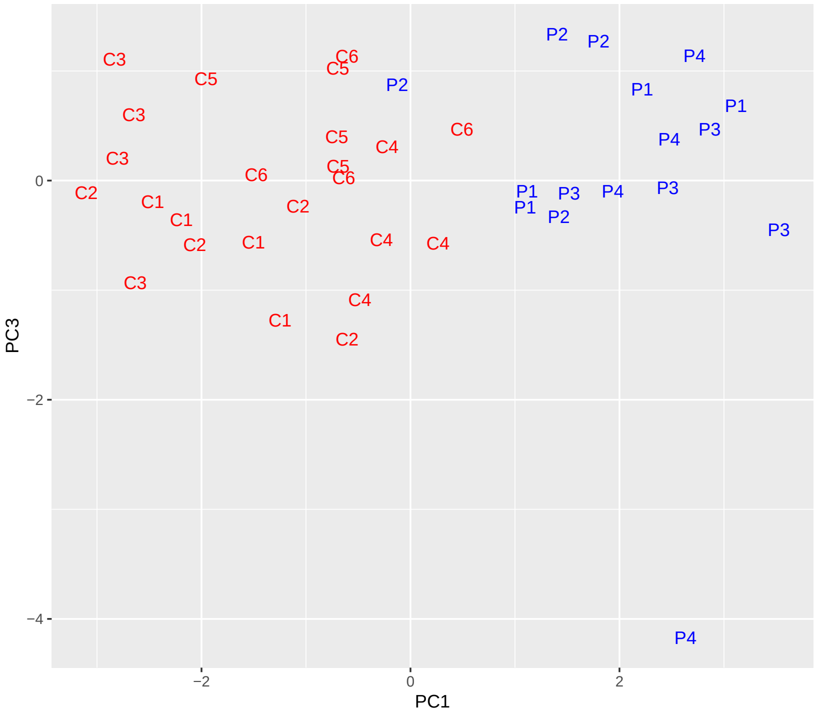


Figure B-6: PCA 1 and PCA 3 of the standardized data [-]

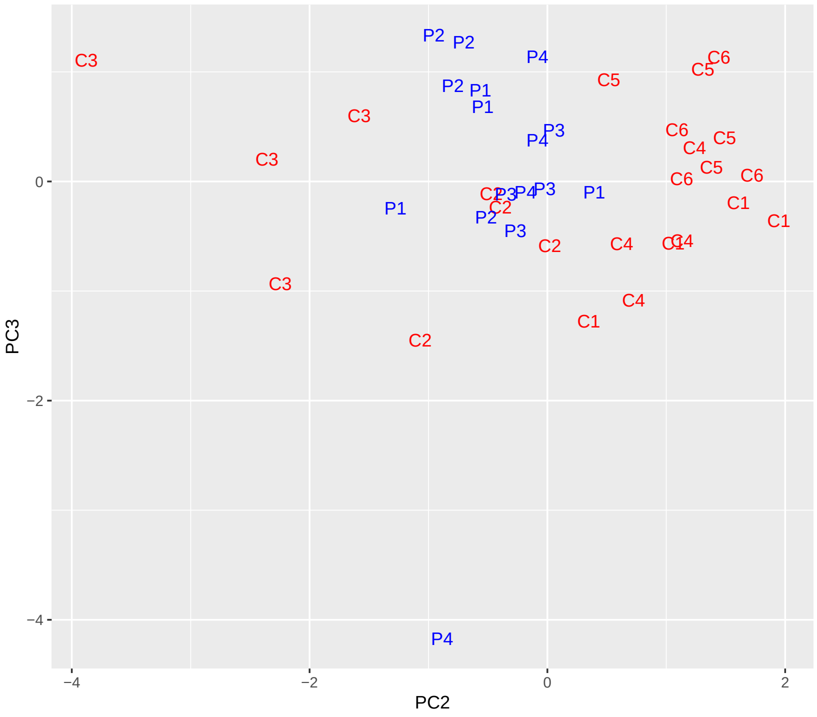


Figure B-7: PCA 2 and PCA 3 of the standardized data [-]

Appendix C

Different Scenario

The dataset consists of data from four different runs solved by three different ATCos. All participants completed a multiple day extensive ATC-course and/or had worked as a researcher in the ATC field (the ATC course/research group) (Van Rooijen, 2018).

The experiment was performed in a medium-fidelity ATC simulation tool called *Sector X* (Van Rooijen, 2018). A sector was used which was comparable to the AMS ACC South Sector. Participant were able to provide commands to the aircraft using the separate command window by means of clicking on buttons using a computer mouse. The traffic was controlled by clicking on the aircraft and then giving a command using the command window. The aircraft were separated by giving heading and speed commands.

The participant had to follow the following specific instructions (Van Rooijen, 2018):

- Loss-of-Separation should be avoided.
- Aim to guide the aircraft to their exit waypoint as efficiently as possible.

Two different scenarios were used in this experiment. Each ATCo performed the scenario twice: one time with SSD and one time without SSD. The scenario had constant sector settings but had different traffic. Each scenario contained 10 conflict pairs caused by 20 aircraft which needed to be solved.

The obtained data consist of two files. One file contains the given commands to the aircraft from the command window, including a timestamp. The other file contains the data from the simulation window. This file includes, per logpoint, the aircraft position, (commanded) heading and (commanded) speed. A logpoint was recorded every 5 seconds during the experiment.

Giving level commands was not part of this experiment and all aircraft in the scenarios were on the same flight level (Van Rooijen, 2018). Therefore, all metrics and measures that incorporate the use of a variable flight level were excluded for the comparison. With the altered total measure set the selection process was run again to find the best set of measures.

Table C-1: Set of measures, without flight level measures, that best describes the expertise level of the ATCos from the dataset of Somers

M1	Total given number of commands
M2	Ratio between the number of given DCT commands and the total number of given DCT, HDG and SPD commands
M3	Ratio between the number of given HDG commands and the total number of given DCT, HDG and SPD commands
M4	Ratio between the total sum of squared track miles when a heading command is given and the total sum of squared track miles when a heading or speed command is given
M5	Ratio between the total sum of squared track miles when a speed command is given and the total sum of squared track miles when a heading or speed command is given
M6	Minimum value of all the occupied SSD Areas
M7	Standard deviation of the aircraft time in sector
M8	Mean of the average TCPA per logpoint
M9	Maximum of the average TLOS per logpoint
M10	Standard deviation of the relative distance between aircraft per logpoint

Table C-1 shows the set of measures, without flight level measures, that best describes the expertise level of the ATCos from the dataset of Somers.

The data of the measures from Table C-1 are extracted from the data from Van Rooijen. The extracted data is standardized with the mean and standard deviation of the measures from the data from Somers. Only then can the data be clustered together to get the results shown in Figure C-1.

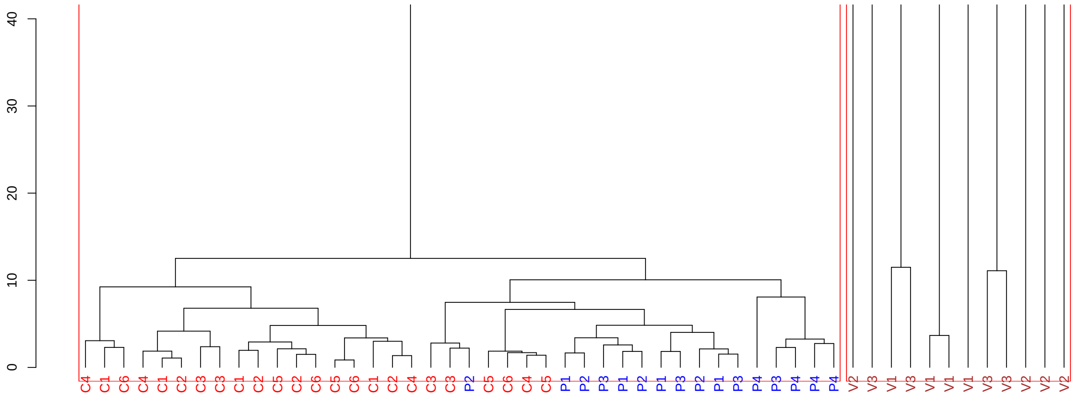


Figure C-1: Dendrogram when adding a test set containing data from a different experiment. The data from Somers is represented by *C* and *P* participants. The data from Van Rooijen is represented by *V* participants

From the figure it can be seen that the data from Van Rooijen is very distant from the data from Somers. The measures used in Table C-1 are scenario dependent and can therefore only be used to compare data from the same experiment.

Appendix D

Software Architecture

This appendix chapter shows the software architecture used to obtain the best set of measures that describes the expertise level. Figure D-1 shows the steps in preparing, processing and clustering the data. To achieve this, a combination of *Python* and *R* is used.

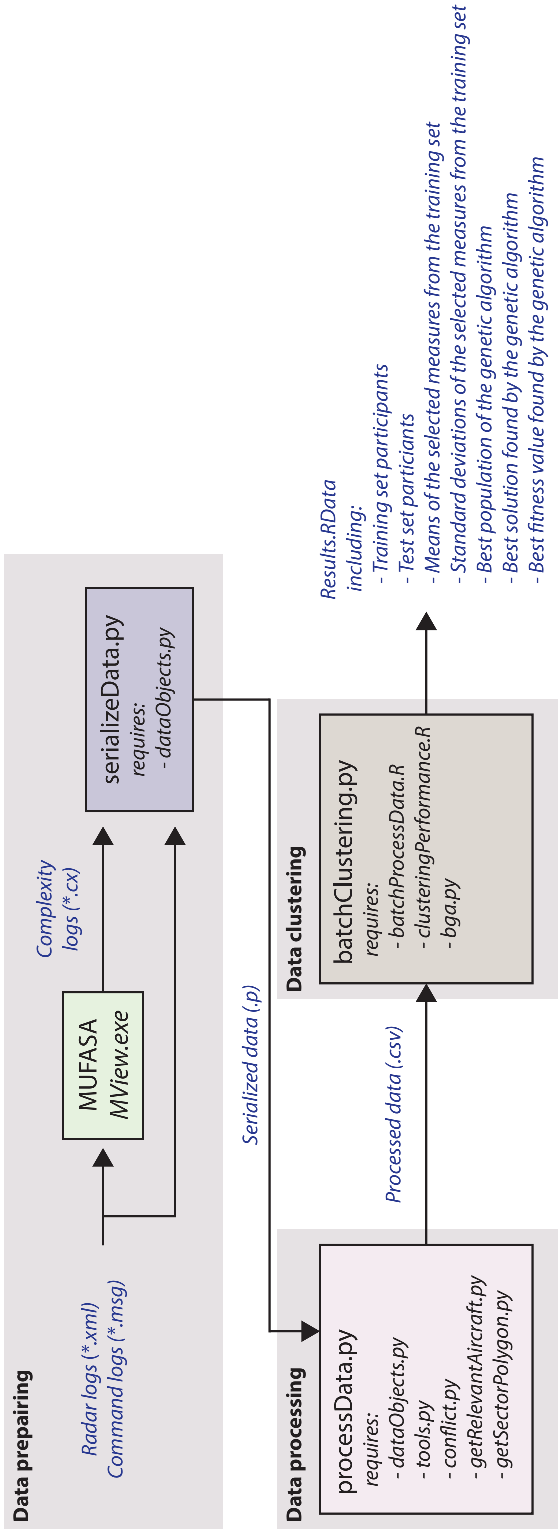


Figure D-1: The software architecture used to find the best set of measures that describes the expertise level

Bibliography

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When Is “Nearest Neighbor” Meaningful? [Conference Paper]. In C. Beeri & P. Buneman (Eds.), *Database Theory — ICDT’99* (p. 217-235). Heidelberg, Germany: Springer.

Boccignone, G., Ferraro, M., Crespi, S., Robino, C., & De’Sperati, C. (2014). Detecting expert’s eye using a multiple-kernel Relevance Vector Machine [Journal Article]. *Journal of Eye Movement Research*, *7*(2).

D’Arcy, J.-F., & Della Rocco, P. S. (2001). *Air Traffic Control Specialist Decision Making and Strategic Planning - A Field Survey* (Technical Report). Federal Aviation Administration, Atlantic City, NJ.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th ed.) [Book]. Chichester, UK: Wiley.

Flanagan, B., Hirokawa, S., Kaneko, E., & Izumi, E. (2017). Classification of Speaking Proficiency Level by Machine Learning and Feature Selection [Conference Paper]. In *1st International Symposium on Emerging Technologies for Education, SETE 2016 Held in Conjunction with ICWL 2016* (Vol. 10108 LNCS, p. 677-682). Rome, Italy: Springer.

Fothergill, S., & Neal, A. (2013). Conflict-resolution heuristics for en route air traffic management [Conference Paper]. In *57th human factors and ergonomics society annual meeting - 2013, hfes 2013* (p. 71-75). San Diego, CA: Proceedings of the Human Factors and Ergonomics Society Annual Meeting 57.

Hilburn, B. (2004). *Cognitive Complexity in Air Traffic Control: A Literature Review* (Technical Report). EUROCONTROL Experimental Centre, Brétigny-sur-Orge, France.

ICAO. (2016). *Doc 4444, Procedures for Air Navigation Services - Air Traffic Management* (Technical Report). ICAO, Quebec, Canada.

Janssen, P., Walther, C., & Lüdeke, M. K. B. (2012). *Cluster Analysis to Understand Socio-Ecological Systems: A Guideline* (PIK Report No. 126). Potsdam-Institut für Klimafolgenforschung, Potsdam, Germany.

Kallus, K., Damme, D. van, & Dittmann, A. (1999). *Integrated Task and Job Analysis of Air Traffic Controllers – Phase 2: Task Analysis of En-route Controllers* (Technical Report). EUROCONTROL, Brussels, Belgium.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis* (1st ed.) [Book]. New York, NY: Wiley.

Kirwan, B., & Flynn, M. (2001). *Identification of Air Traffic Controller Conflict Resolution Strategies for the CORA (Conflict Resolution Assistant) Project* (Technical Report). EUROCONTROL Experimental Centre, Brétigny, France.

Loft, S., Sanderson, P. M., Neal, A., & Mooij, M. (2007). Modeling and Predicting Mental Workload in En Route Air Traffic Control: Critical Review and Broader Implications [Journal Article]. *Human Factors*, *49*(3), 376-399.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval* (1st ed.) [Book]. Cambridge, UK: Cambridge University Press.

Mercado Velasco, G. A., Mulder, M., & Paassen, M. M. van. (2010). Analysis of Air Traffic Controller Workload Reduction Based on the Solution Space for the Merging Task [Conference Paper]. In *AIAA Guidance, Navigation, and Control Conference.* Toronto, Canada: American Institute of Aeronautics and Astronautics (AIAA).

Negnevitsky, M. (2011). *Artificial Intelligence: A Guide to Intelligent Systems* (3rd ed.) [Book]. Harlow, UK: Addison Wesley.

Oprins, E., Burggraaff, E., & Weerdenburg, H. van. (2006). Design of a Competence-Based Assessment System for Air Traffic Control Training [Journal Article]. *The International Journal of Aviation Psychology*, *16*(3), 297-320.

Oprins, E., & Schuver, M. (2003). *Competentiegericht opleiden en beoordelen bij LVNL (Competence-based training and assessment at LVNL)* (Newsletter Article No. 6). Human Factors Advisory Group, Schiphol, The Netherlands.

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence : A Modern Approach* (3rd ed.) [Book]. Upper Saddle River, NJ: Prentice Hall.

Schuver-van Blanken, M. J., Huisman, H., & Roerdink, M. I. (2010). The ATC Cognitive Process and Operational Situation Model - A model for analysing cognitive complexity in ATC [Conference Paper]. In *29th EAAP Conference.* Budapest, Hungary.

Schuver-van Blanken, M. J., & Merriënboer, J. G. van. (2012). Air Traffic Controller Strategies in Operational Disturbances - An exploratory study in air traffic control [Conference Paper]. In *30th EAAP Conference.* Sardinia, Italy.

Schuver-van Blanken, M. J., & Roerdink, M. I. (2013). Clarifying Cognitive Complexity and Controller Strategies in Disturbed Inbound Peak ATC Operations [Conference Paper]. In *17th International Symposium on Aviation Psychology.* Dayton, OH.

Somers, V. L. J. (2017). *3D Solution Space-based Prediction of Air Traffic Control Workload.* Unpublished M.Sc. Thesis, Faculty of Aerospace Engineering, Delft University of Technology.

Stańczyk, U., & Jain, L. C. (2015). *Feature Selection for Data and Pattern Recognition* (1st ed.) [Book]. Heidelberg, Germany: Springer.

Van Rooijen, S. J. (2018). *Towards Personalized Automation for Air Traffic Control using Convolutional Neural Networks*. Unpublished M.Sc. Thesis, Faculty of Aerospace Engineering, Delft University of Technology.

Watson, R. A. (2014). Use of a Machine Learning Algorithm to Classify Expertise: Analysis of Hand Motion Patterns During a Simulated Surgical Task [Journal Article]. *Academic Medicine*, *89*(8), 1163-1167.

Xu, R., & Wunsch Ii, D. (2005). Survey of Clustering Algorithms [Journal Article]. *IEEE Transactions on Neural Networks*, *16*(3), 645-678.